A Fast, Easy, and Stable Approach to Dynamic Analysis

Daniel Kent[*] James Wilson[†] Skyler J. Cranmer[‡]

Word Count: 6,870

Abstract

Across the social sciences scholars regularly pool effects over substantial periods of time, a practice that produces faulty inferences if the underlying data generating process is dynamic. To help researchers better perform principled dynamic analysis, we draw on the literature on statistical process monitoring and develop a method for detecting temporal shifts in model parameters. This technique performs as well as Bayesian changepoint analysis in simulation, but with minimal assumptions and fewer conceptual barriers. Indeed, applying the method appropriately only requires an understanding of a sample's mean and variance. We also demonstrate the method's utility by applying it to a popular study on the relationship between alliance types and the initiation of militarized interstate disputes. The example illustrates how the technique can help researchers make inferences about where changes occur in dynamic relationships and ask important theoretical questions about the causes and consequences of such changes.

[*] Corresponding author: kent.249@osu.edu

[†] jdwilson4@usfca.edu

[‡] cranmer.12@osu.edu

Introduction

Political scientists know Simpson's paradox – the phenomenon of clear trends appearing within groups of data but vanishing when the data are pooled – well and it is the impetus behind a large body of work in hierarchical/multilevel modeling. Yet few consider that a variant of Simpson's paradox can occur with effects that vary over time, regardless of group-level variation: time trends in effect size, particularly the non-linear sort, can be masked by pooling over time periods that display substantial temporal heterogeneity in effects. Here, we provide a fast, simple, and stable procedure for evaluating when and where effects are genuinely dynamic as opposed to noisy but stable.

Many important social and political processes – ranging from international conflict to democratization and from voting behavior to public policy – manifest over time. Since these are inherently longitudinal processes, it is natural that their temporal dynamics be a major focus for researchers. Yet while prominent studies in these areas do describe the temporal aspect of the data, few consider that the effects of interest – those measuring the *relationship* between the predictor and the outcome of interest – might vary over time.

Indeed, the overwhelming majority of empirical studies in Political Science assume the effects of interest are stable over time. In time series analysis, the primary emphasis is on modeling a process as a function of its history (e.g. lagged outcome variables) or on detrending a series (e.g. by modeling a differenced outcome variable) (Box-Steffensmeier et al., 2014; Enders, 2004; Huckfeldt et al., 1982; McCleary et al., 1980). When extended to time series cross-sectional analysis, the inclusion of fixed/random effects (often in addition to detrending or lagged outcome inclusion) absorbs variation across temporal and/or observational units (Angrist and Pischke, 2009; Beck, 2008; Bell and Jones, 2015; Clark and Linzer, 2015; Gelman and Hill, 2007; Green et al., 2001; Plümper and Troeger, 2007). While the methods just described account for temporal processes, they implicitly assume that the relationships between predictors and response remain constant through time. For example, dynamic regression models that include a lagged dependent variable along with predictors (often referred to as panel models) do not allow for the fact that the temporal dependency of any of the predictors might change their relationship with the outcome variable during the period of observation. These families of models assume homogeneous effects over time.

The tacit assumption of temporal stability of effects is actually quite strong, often unrealistic, and can result in faulty inference when applied inappropriately. The researcher must ask whether it really makes sense to have a single coefficient reflect the relationship between predictor and outcome over a period of time. If the process is stable, this is appropriate. If the process is dynamic, it is not. For example, Cranmer et al. (2014) find substantial effect heterogeneity in their study of international sanctions networks from 1972 to 2000.Reducing a dynamic process to a single point estimate risks averaging out important variation in the process. If an effect is positive for half the period of observation and negative for the other half (with the same magnitude), assuming a static effect and pooling over the whole series may lead to an average pooled effect of zero, which leads to false inference and interpretation. In instances where the effect spends more time on one side of zero than the other, or with greater magnitude on one side than the other, the result of falsely assuming a static process will be a misleading positive (negative) coefficient.

Moreover, assuming static effects forfeits the opportunity to learn about, and even test hypotheses about, the dynamic nature of socio-political phenomena. Though we are not the first to express concern over the lack of dynamic modeling in political science (Beck et al., 1998; Beck, 2001, 2008; Box-Steffensmeier and Jones, 2004; Carter and Signorino, 2010; De Boef and Keele, 2008; Gelman and Hill, 2007; Gill, 2014; Golub, 2008; Jenke and Gelpi, 2016; King, 1998; Mitchell et al., 1999; Nieman, 2016; Park, 2012; Wawro and Katznelson, 2014; Zorn, 2001), our aim here is to present a comprehensive, intuitive, and easy-to-implement solution. This stands in contrast to popular Bayesian methods, which tend to be complicated, computationally demanding, and time-intensive.

We introduce and implement a statistical process monitoring technique for the family of generalized linear models that dominate statistical analysis in the social sciences, and is easily extended to more complicated models (e.g. network models). Statistical process monitoring (SPM) is a statistical technique that provides a methodology for the real-time Shewhart chart of any statistic that fluctuates through time. SPM is used to flag anomalous behavior in such a characteristic by distinguishing unusual variation from typical variation in an ordered sequence of observations. Our approach, at its most basic level, functions as a simple test for the presence of effect heterogeneity. Indeed, we also develop a statistical test for temporal effect stability. If evidence of heterogeneity is detected by this test, our technique then allows the researcher to explore the effect's dynamicism and generate new knowledge about the process. The techniques that we introduce here are implemented in a free and easy-to-use companion software in the R language.

The implications of our work are substantial. The empirical literature in political science rarely considers the problem of effect stability. A survey of the *American Political Science Review* reveals that in 2015 alone, 10 out of its 48 total articles employed longitudinal data and estimated models with pooled effects. Though hardly a comprehensive survey of the field, the implication is that a substantial portion of the articles in our discipline potentially suffer from substantial unmodeled temporal effect heterogeneity. In the replication study we conduct below to illustrate our technique we estimate that the relationship changes five times in magnitude, whereas the original paper only includes a single pooled analysis spanning over 100 years.[1] Fortunately, the relationship does not change in direction, but its magnitude is highly variant over time. In this context, testing for and modeling temporal heterogeneity not only brings valuable context to the analysis, it also presents an opportunity to learn about a substantively important dynamic process.

[1] In our supplement we consider three other well-cited articles from various subfields of Political Science, all of which include significant unmodeled dynamic effects.

Why Worry About Dynamic Effects

The vast majority of statistical models applied in political science research are based on the generalized linear model (GLM; e.g. linear, logistic, and Poisson regression) and these GLMs are frequently applied to temporal data, either in the form of time series or time series cross-sectional specifications. Political scientists have come to take time increasingly seriously via lagged outcome variables (Angrist and Pischke, 2009; Box-Steffensmeier et al., 2014; Franzese Jr and Hays, 2007; Keele and Kelly, 2005; Plümper and Troeger, 2007), detrended time series (Beck and Katz, 1995; Box-Steffensmeier et al., 2014), fixed/random effects (Angrist and Pischke, 2009; Beck, 2008; Bell and Jones, 2015; Clark and Linzer, 2015; Gelman and Hill, 2007; Green et al., 2001; Plümper and Troeger, 2007), and splines (Hastie and Tibshirani, 1990; Beck et al., 1998; Gelman and Hill, 2007). Yet the temporal models with which the quantitative researcher is now familiar are not dynamic models in that they do not allow the *effect itself* to vary over time. If the effect – the relationship between predictor and outcome as captured by the estimated coefficient – changes during the period of observation, then the researcher risks faulty inference by failing to account for these dynamics.

Modeling time is not the same as having dynamic effects in a model. Consider the simple example of a model with a lagged outcome variable as recommended by Beck and Katz (1996). This model predicts the outcome based on the previous observation of the outcome (and can naturally accommodate other predictors as well). But the relationship between the lagged outcome and the current outcome, as well as the relationship between any predictor and the outcome, is assumed to be static; the model produces a single coefficient describing this relationship over the entire period of observation. Dynamic models account for the fact that the relationship between the outcome and any of its predictors may change over time.

It is difficult to think of relationships in political science where the assumption of temporal stability is not worth questioning. Has the effect of wealth on democratization remained static over the last 200 years? Has the effect of partisanship in Congress changed since the Congress was founded, or even in the last 30 years? Why then are dynamic effect models so rare in political science when temporal data are so plentiful? We believe this is caused in part by the fact that static effect models are easier, both theoretically and methodologically. Dynamic models can require the researcher to specify a more detailed theory about how effects shift over time. For instance, a conflict researcher studying deterrence might need to develop clear expectations about how deterrent relationships change when states develop nuclear weapons, as it seems likely that a deterrent effect would change after such a development. Empirically, dynamic models require more careful and nuanced interpretations than static models, placing an increased explanatory burden on the researcher when presenting results. Through the companion software to this manuscript, we significantly reduce the complexity of specifying and interpreting dynamic models.

Failure to account for dynamic effects, when they exist in the data, comes with two costs: the researcher risks faulty inference with static effects and forfeits the ability to learn about substantively meaningful variation in the process of interest. Faulty inference can occur in several ways when dynamic effects go unmodeled. First, type II errors (false negatives) are possible. Imagine that an effect spends about half the period of observation below zero at a magnitude of about one, then the other half of the period of observation above zero with similar magnitude. A temporally pooled model will not detect this shift in the relationship and will average the effect out to zero, leading the researcher to erroneously conclude that there is no relationship between the predictor and the outcome. Type I errors (false positives) are also possible. Consider an effect that is positive though not statistically reliable over most of the period of observation, but with a few "spikes" that are positive and statistically reliable. A model with temporally pooled effects will lead the researcher to falsely conclude that there is a non-null relationship over the entire period of observation, whereas the ground truth is that the effect is null for all but a few time points.

A third type of faulty inference can occur, which is the incorrect conclusion that effects are stable over time. A model with temporally pooled coefficients will, by design, be unable to detect shifts in the relationship between predictor and outcome. This means that substantively important variation can go undetected and lead the researcher to falsely conclude that a single, static, coefficient represents the relationship well. This last faulty inference is a type of specification error. For example, in figure 1 we apply our technique to Leeds' (2003) study on the relationship between alliance types and militarized interstate disputes (MIDs). In this figure we focus on the relationship between defensive alliances and the outbreak of a MID and estimate a separate model for every temporal interval identified by our technique, as opposed to a single pooled model. As the figure shows, the relationship is highly dynamic, which one would expect, given the temporal length of the study.

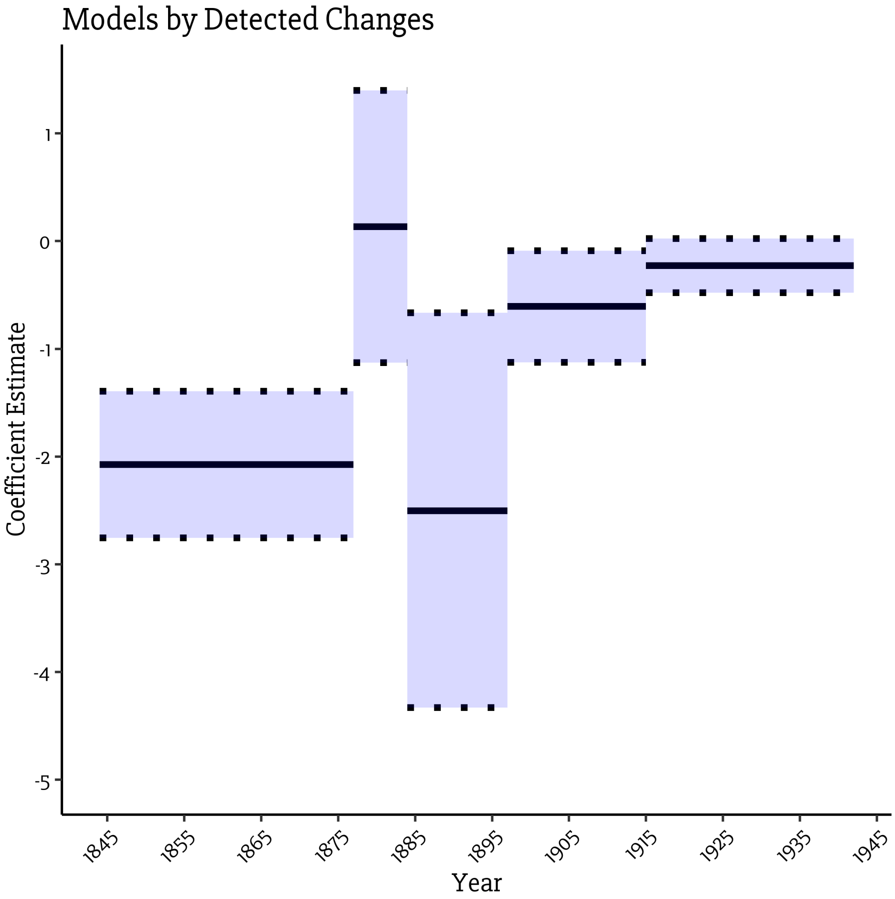

Figure 1: Replication and extension of Leeds (2003), where the y-axis is the coefficient estimate for the association between defensive alliances and the initiation of militarized interstate disputes. Each section represents a separate model. The temporal bounds for each model are recommended by our proposed technique.

Our claim is not that temporal heterogeneity has gone unnoticed. Indeed, there have been several works in Political Science that do account for response heterogeneity through time. For example, Brandt and Sandler (2009), Freeman et al. (2000), Nieman (2016), Park (2011a), Park (2011b), Park (2012), Scheve and Stasavage (2009), and Spirling (2007) all employ change-point models to detect response heterogeneity in applications to economic development, exchange rates, terrorism, wartime casualties, and more.

Our proposed strategy builds on the work of Wawro and Katznelson (2014), who embrace techniques for detecting parameter heterogeneity as a means of bridging the quantitative-qualitative divide, particularly in the field of American Political Development. They present Bayesian change point and structured additive regression (STAR) models as methods for detecting effect heterogeneity. A major difference between our monitoring technique and the STAR model is that our technique is designed to capture time points where the estimated relationship has changed, whereas the STAR model solely allows for its estimated relationship to vary over time. In this regard, if a researcher not only wants temporal flexibility, but wishes to ask where meaningful changes have occurred and why, then our method provides significant additional leverage. Though a STAR model may be used in conjunction with Bayesian changepoint analysis to detect where changes occur, the latter carries the aforementioned conceptual and computational barriers. As we discuss below, our technique performs as well as Bayesian changepoint analysis in simulation with fewer conceptual and computational barriers. In fact, in the presence of highly dynamic relationships, our technique returns more accurate estimates.

Taking a different approach, in their oft-used methods textbook for social scientists, Gelman and Hill (2007, pp. 73) make a similar suggestion to ours in their discussion of a "secret weapon" – separately estimating regression coefficients for each group or time unit and then plotting the estimates out. Cranmer et al. (2014) develop this idea further in their study of the international sanctions network by plotting effects annually and then using a combination of change point models and nonparametric smoothers to create smoothed estimates over time. Jenke and Gelpi (2016) apply this same basic concept to the study of conflict, replicating Bennett and Stam's (2009) model of interstate conflict and breaking it into three time periods: pre-WWI, the interwar years, and post-WWII. Using out of sample prediction to compare the pooled model to their separate temporal-specific models, they find support for taking historical eras into account. Making similar arguments in the study of international politics: Braumoeller (2013) analyzes three discrete international systems, Gleditsch and Ward (2000) argue that the effect of democratization on war depends both on regional and temporal contexts, Mitchell et al. (1999) find that the relationship between democracy and war varies with systemic configurations, and Nieman (2016) finds that the relationship between trade and conflict changes significantly with the start of World War II.

Our discipline requires techniques that will allow us to easily identify when effects are dynamic during a given period of observation and to explore those dynamics when they are present. This is not as easy as it may sound; the principal challenge lies in differentiating minor stochastic fluctuations in key effects from real changes in the data generating process. In the following sections, we introduce a method that accomplishes just this.

Testing and Monitoring Generalized Linear Models

In a panel study, one observes multivariate measurements on n samples through time. These measurements can be represented by the sequence $\{(\mathbf{X}_t, Y_t),\ t = 1, \ldots, T\}$, where $Y_t \in \mathbb{R}^n$ is the outcome response variable at time t, and $\mathbf{X}_t$ is an $n \times p$ matrix of p predictors at time t. In many cases, the goal of a panel study is to estimate the relationship between Y_t and $\mathbf{X}_t$. One does this by treating Y_t as a random variable whose mean μ_t relates to the predictors $\mathbf{X}_t$ in the form of some generalized linear model described as follows

$$g(\mu_t) = \mathbf{X}_t \boldsymbol{\beta}, \quad t = 1, \ldots, T. \tag{1}$$

Here, $g(\cdot)$ is an appropriately chosen link function that maps μ_t to a linear function of the predictors $\mathbf{X}_t$. Note that an intercept for fixed effects can be incorporated in Equation (1) by setting the first column of $\mathbf{X}_t$ to be a column of 1s for all t. Common choices of $g(\cdot)$ include the identity, logit, and natural log function, corresponding to multivariate linear, logistic, and

Poisson regression, respectively.

As specified, the coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ quantify the *average* or *pooled* effect of $\mathbf{X}_t$ on the mean of Y_t. It follows that model (1) requires a static relationship between $\mathbf{X}_t$ and Y_t, one that does not vary through time. In many applications, this assumption simply does not hold as the relationship between $\mathbf{X}_t$ and Y_t for at least one key variable is dynamic. For example, in the Leeds replication that we previously described, the relationship between defensive alliances and the initiation of militarized interstate disputes as shown in Figure 1 changes significantly over time, reflecting variation in the frequency of conflicts over time and variation in the composition of international security regimes. In studies of this nature, one should test for dynamics in the effect of each covariate of interest before applying a pooled model like that in equation (1). In light of this, we provide diagnostics to answer two questions about fitting a pooled model to panel data:

1. Is the effect $\boldsymbol{\beta}$ constant through time?

2. If $\boldsymbol{\beta}$ is not constant, then at which time points is the effect significantly different than what we would expect under a valid pooled model?

We address these two questions in turn.

Testing the Pooled Model for Dynamicism

Our first goal is to determine whether or not the pooled model in equation (1) is valid for an observed panel $\{(\mathbf{X}_t, Y_t),\ t = 1, \dots, T\}$. That is, we quantify whether the coefficient $\boldsymbol{\beta}$ remains constant through time. In the case that $\boldsymbol{\beta}$ does not remain constant, and the pooled model is not appropriate, one should instead fit the following alternative model

$$g(\mu_t) = \mathbf{X}_t \boldsymbol{\beta}_t, \quad t = 1, \dots, T. \tag{2}$$

Model (2) is commonly known as a varying-coefficient model for longitudinal data (Hastie and Tibshirani, 1993; Hoover et al., 1998). Varying-coefficient models have been well studied in statistics (see Fan and Zhang (2008) for a review). These models provide a time-varying alternative to the otherwise static pooled model considered in (1). Indeed, model (1) is a special case of model (2) in which $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}$ for all t. The nested relationship of these two alternative models suggests a principled manner to formally test the validity of the pooled model through the use of a generalized likelihood ratio test which tests the null hypothesis that the pooled model is appropriate. In what follows, we first describe the hypothesis testing framework and then provide details about how to calculate the generalized likelihood ratio statistic to formally test the null hypothesis through the use of a bootstrap sample from the observed panel.

Let $\boldsymbol{\beta}_t = (\beta_1(t), \ldots, \beta_p(t))$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Assessing the homogeneity of model (1) can be formulated as a nested-model hypothesis test, where the null hypothesis reflects the pooled model

$$H_0 : \beta_k(t) = \beta_k \qquad k = 1, \ldots, p, \quad t = 1, \ldots, T. \tag{3}$$

We can test (3) using the generalized likelihood ratio test developed in Fan and Zhang (2000) and Fan et al. (2001). Let $\widehat{\boldsymbol{\beta}}$ and $\{\widehat{\boldsymbol{\beta}}_t, t = 1, \ldots, T\}$ represent the maximum likelihood estimators of the coefficients from models (1) and (2), respectively. Furthermore, let $\ell(\cdot \mid Y_t)$ represent the joint log-likelihood function for the tth observation in the panel. The generalized likelihood ratio statistic is the difference between the log-likelihood functions under the alternative and null hypothesis:

$$\Gamma_0 = \sum_{t=1}^{T} \left[\ell(g^{-1}(\mathbf{X}_t \widehat{\boldsymbol{\beta}}_t \mid Y_t)) - \ell(g^{-1}(\mathbf{X}_t \widehat{\boldsymbol{\beta}} \mid Y_t)) \right]. \tag{4}$$

As observed in Fan et al. (2001) (Section 4.2), the Wilk's phenomenon implies that as $T \to \infty$, Γ_0 is asymptotically Normal and does not depend on the estimates $\widehat{\boldsymbol{\beta}}$. In practice, one can directly use this asymptotic distribution to test the validity of model (1). However, in the present context we employ a bootstrap procedure to test H_0 on the basis of its improved power over the former approximate distribution when T is small (De Brabanter et al., 2006; Lee et al., 2017). The bootstrap testing procedure is a simulation-based technique, with a user-selected number of simulations N, that consists of the following three steps.

1. For $n = 1, \ldots, N$, generate a sample $y^{(n)} = \{y_t^{(n)} : t = 1, \ldots, T\}$ under the estimated null pooled model in equation (1).
2. For each bootstrap sample $y^{(n)}$, estimate $\widehat{\boldsymbol{\beta}}^{(n)}$ and $\widehat{\boldsymbol{\beta}}_1^{(n)}, \ldots, \widehat{\boldsymbol{\beta}}_T^{(n)}$ under models (1) and (2), respectively, using the observed data $\mathbf{X}_1, \ldots, \mathbf{X}_T$.
3. Calculate the test statistic $\Gamma^{(n)}$ for each bootstrap sample:

$$\Gamma^{(n)} = \sum_{t=1}^{T} \left[\ell(g^{-1}(\mathbf{X}_t \widehat{\boldsymbol{\beta}}_t^{(n)} \mid y_t^{(n)})) - \ell(g^{-1}(\mathbf{X}_t \widehat{\boldsymbol{\beta}}^{(n)} \mid y_t^{(n)})) \right].$$

We test the null hypothesis by comparing the observed statistic Γ_0 with the collection of bootstrapped statistics $(\Gamma^{(1)}, \ldots, \Gamma^{(N)})$. The p-value to test H_0 is the proportion of bootstrap test statistic values that exceed the observed test statistic Γ_0. In particular, let $\mathbb{I}(A)$ be the indicator function that takes the value 1 if A is true and 0 otherwise. Then the p-value for testing the null hypothesis in Equation (3) is given by $p = N^{-1} \sum_{n=1}^{N} \mathbb{I}\left(\Gamma_0 < \Gamma^{(n)}\right)$. Small values of p suggest that Γ_0 is smaller than we would expect under H_0 and therefore provide evidence against the pooled model. In general, one selects a level $\alpha \in (0, 1)$ and decides to reject H_0 in favor of the varying coefficient model whenever $p < \alpha$. Our free and easy-to-use companion software performs this bootstrap test for any generalized linear model implemented in R's `glm` function. See the supplement to this manuscript for a software example.

Monitoring Panel Data with Moving-Window Control Charts

If the panel $\{(\mathbf{X}_t, Y_t),\ t = 1, \ldots, T\}$ is deemed heterogeneous according to the rejection of the null hypothesis in Equation (3), then it may be of interest to identify at which point(s) the coefficients are statistically different than the pooled estimates or from prior estimates. One can readily adapt statistical process monitoring (SPM) on the estimated coefficients $\{\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_T\}$ to identify such changes. We propose a SPM-influenced technique as a favorable alternative to Bayesian changepoint analysis, because of its relative computational and conceptual accessibility. We monitor the maximum likelihood estimators of the coefficients $\widehat{\boldsymbol{\beta}}_t$, which quantify the effects of $\mathbf{X}_t$ on the response Y_t.

Statistical process monitoring is a widely studied branch of industrial statistics that provides a methodology for the real-time monitoring of a stochastic process (Montgomery and Keats, 1991; MacGregor and Kourti, 1995; Oakland, 2007; Woodall et al., 2017). The aim of SPM is to identify anomalous behavior in a process by distinguishing unusual variation from typical variation in an ordered sequence of observations. Stemming from applications in industrial manufacturing and public health surveillance, SPM has a rich history for which many techniques have been developed (see e.g., Woodall and Montgomery (1999), Woodall and Montgomery (2014), Benneyan et al. (2003) for reviews of methods and applications). Here, we briefly describe SPM to motivate its adaptation to the monitoring of coefficients over time in social and political panel studies.

Our goal is to prospectively monitor the panel $\{(\mathbf{X}_t, Y_t),\ t = 1, \ldots, T\}$ to detect heterogeneous effects through time. To perform surveillance, one first specifies a statistic S_t, or more generally a vector of statistics $\boldsymbol{S}_t$, that provides a summary of the data ($\mathbf{X}_t$, Y_t). The choice of S_t is flexible. In our case of a fitted GLM, the vector of statistics $\boldsymbol{S}_t$ can be the estimated coefficients of the model at time t.

SPM seeks the real-time identification of unusually large or small values of S_t through the use of a *control chart* – a time series plot of S_t constructed with *control limits* that indicate boundaries of typical behavior. S_t is considered anomalous, or signalled, if it deviates significantly from what observations suggest is typical. As exemplified in our motivating example, the observed panel may exhibit multiple changes through time. To identify these changes, we develop a moving-window approach to monitor S_t, described as follows. First, one selects a training period m. Then we iteratively follow two steps:

Training. The statistic S_t is calculated for all data $\{(\mathbf{X}_t, Y_t)\}$ with $t \leq m$. A tolerance region $\mathcal{R}(m)$ is constructed based on these values. The upper and lower bounds of this region are referred to as upper and lower control limits, respectively. Variation within these limits defines typical behavior.

Test. Then, for each new data point $(\mathbf{X}_t, Y_t)$, with $t > m$, S_t is calculated, and the data point is deemed "typical" if $S_t \in \mathcal{R}(m)$ and deemed "anomalous" otherwise.

For the first point t^* identified as anomalous, we begin the **Training** process again by setting $t^* + 1$ to 1, namely the first data point in the training set. By re-training once we've identified a change, we can readily identify multiple mean shifts in the panel data. We continue to re-train in this way in the case of any anomalous points in the panel. An example of a moving-window control chart is illustrated in Figure 2.

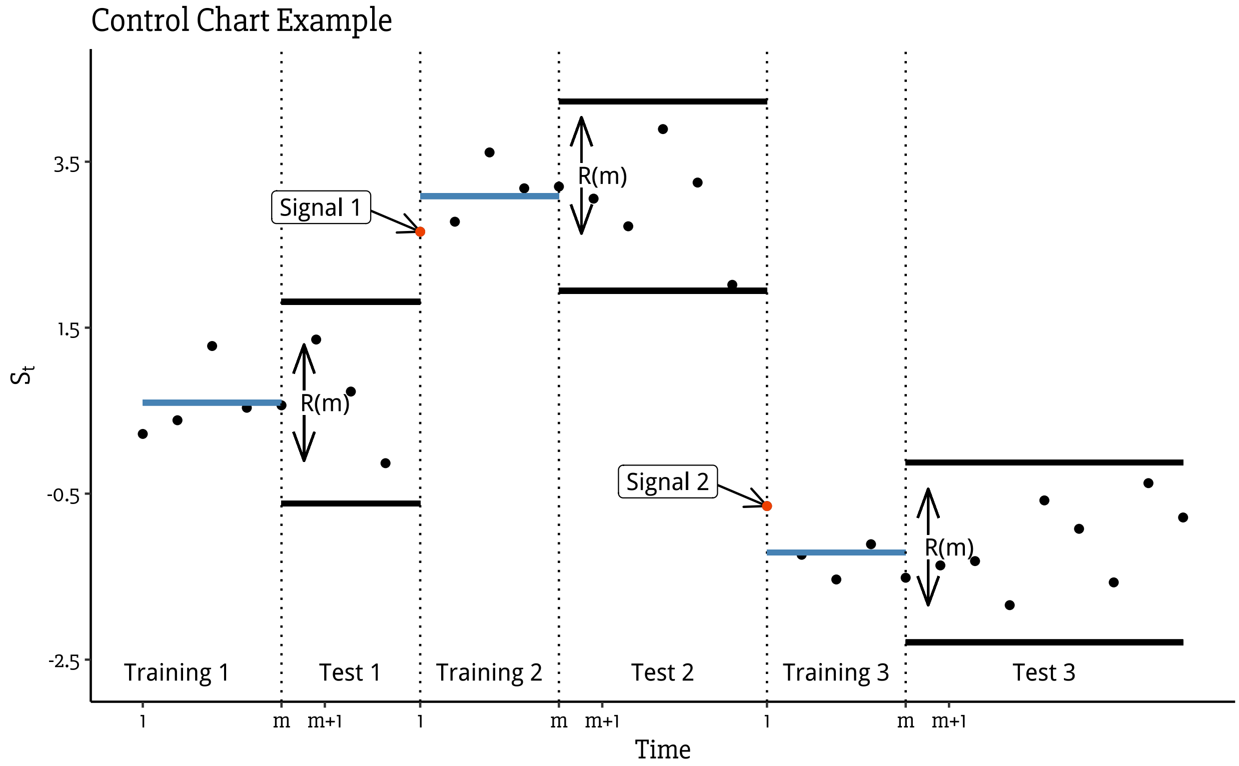

Figure 2: An example of a moving-window control chart. The stochastic process S_t represents the quantity of interest from multivariate data. In each training phase, a tolerance region $\mathcal{R}(m)$ is calculated based on the observed values of S_t for times $t = 1, \ldots, m$. In each subsequent test phase, new data is monitored. A new observation t^* is deemed anomalous or signalled if S_{t^*} lies outside the tolerance region $\mathcal{R}(m)$.

To identify effect heterogeneity in a panel GLM model, we apply SPM to detect significant changes in $\widehat{\boldsymbol{\beta}}_t$. Such alterations correspond to marked changes in the underlying generative process of the data, and suggest heterogeneity in the effect of $\mathbf{X}_t$ on Y_t. For each of the coefficients that we estimate, we use a Shewhart control chart for individual outcomes (Montgomery, 2013) to determine what values indicate a significant change. The Shewhart control chart is described intuitively as follows. For simplicity, suppose that S_t is a one-dimensional statistic at time t, and let m be the size of the training set. For data points in the test set, the Shewhart control chart for individuals outcomes signals a change in the statistic if S_t lies outside of the control limits $\widehat{\mu} \pm 3\widehat{\sigma}$, where $\widehat{\mu}$ is the sample mean of the m training observations, and $\widehat{\sigma}$ is the moving range estimate for the standard deviation of these m observations given by

$$\widehat{\sigma} = \frac{\sqrt{\pi}}{2(m-1)} \sum_{j=2}^{m} |S_j - S_{j-1}|.$$

The $\pm 3\widehat{\sigma}$ represents the margin of error of prediction for new values of μ_t under normality of S_t. We note that the Shewhart chart is one of many possible choices of a control chart from the SPM literature. We favor the Shewhart chart because it is particularly well suited to capture large sudden changes in S_t. For slow cumulative changes or small changes in the observed sequence, one may instead utilize the cumulative sum or exponentially weighted moving average control charts (see Woodall and Montgomery (1999) for an overview of other possible methods.) Our strategy of monitoring the coefficients of a fitted model is related to the work of Wilson et al. (2016), who monitored the estimated coefficients of parametric statistical network models through time. Their work revealed that one can efficiently monitor changes in longitudinal data by monitoring fitted estimates to a possibly dynamic model.

Monte Carlo Simulation

In order to test the method's performance we conducted a Monte Carlo simulation, extracting the Shewhart chart's performance across each iteration and comparing it to cutting-edge software for changepoint analysis.[2] (Erdman et al., 2007) Our criteria for choosing a comparison method was to find a changepoint package which was: (1) fast, (2) performed well in simulations, and (3) did not require pre-specifying the expected number of changes. With regard to the last condition, while the prominent software for changepoint analysis employed in Political Science tends to return accurate estimates when the pre-specified number of changes is correct, it requires that the user pre-specifies the number of changes ahead of time, which presumably one does not know.[3] On the other hand, bcp is fast, flexible, and requires minimal assumptions. Nonetheless, effective implementation does require an understanding of MCMC sampling and posterior probability distributions.

The steps of each simulation iteration were specified so that data for a dynamic regression model, $y_t = X_t\beta_t$, was generated with a pre-specified number of changes in β. During each iteration a regression for every time period was fit, both methods were applied, and the difference between the true and estimated changes was stored. Across simulations, σ was varied in magnitude when β_t was generated, allowing us to manipulate the degree of noise around each stable time series – creating additional variation in the difficulty of cases for both techniques. Specifically, simulations were conducted as follows:

- The number of changepoints, P, was drawn from a truncated normal distribution.[4]
- P draws from a discrete uniform distribution on 1 to 100 were made and each value (start and finish of the time series) was used as the location of a parameter change.
- Mean values μ_t for $t = 1, \dots, P$ were randomly drawn from a standard normal distribution.
- For $t \in [1, T]$ data was generated according to a linear model: $y_t = X_t\beta_t$ where each β_t was drawn from a normal distribution with mean μ_t and variance 1. Any variation in the

[2] See the bcp R package: `https://cran.r-project.org/web/packages/bcp/bcp.pdf`.

[3] See `MCMCpack`

[4] Sampling from a truncated normal distribution ensured a positive number of parameters were selected. The distribution has a minimum of 1, mean of 1, and standard deviation of 2. Each draw was also rounded to a whole number.

magnitude of each coefficient was a result of being drawn from a probability distribution. After each coefficient change occurred the data generating process remained constant until the next change.

- The Shewhart chart and Bayesian technique were applied to the panel data (X_t, y_t) to estimate the number of parameter changes as well as their location.
- The difference between the actual and estimated number of changes by each method was recorded.

An analysis of the simulation's entire output reveals that while the mean-squared error for both techniques is statistically indistinguishable, there is a bias-variance tradeoff in choosing between the two. Indeed, the Shewhart chart's errors are centered around zero with an expectation of capturing the correct number of changes, whereas changepoint analysis is essentially truncated at zero with a greater tendency for type-II errors and expectation of falling short by one changepoint. With respect to its variance, the Shewhart chart has greater spread in its errors, meaning one risks occasionally missing by a greater magnitude than if they employ changepoint analysis.

Decomposing the simulation's results further, a closer look reveals that as the time series of coefficient estimates grows more dynamic, then the better the Shewhart chart tends to perform relative to changepoint analysis. In Figure 3 we subset the simulation by the number of coefficient changes built into each iteration. The y-axis corresponds to the number of coefficient changes built into an iteration and the x-axis represents our measure of technique performance: the estimated number of coefficient changes minus the actual number of coefficient changes.[5] At each number of changes a boxplot of the summary statistics for both techniques is displayed, where the blue box corresponds to changepoint analysis and the red box to the Shewhart chart. Each boxplot then includes the summary statistics of the technique's

[5] With respect to the estimated number of changes minus the actual number, a negative result means the technique under-estimated the number of changes and a positive result means the technique over-estimated the number of changes.

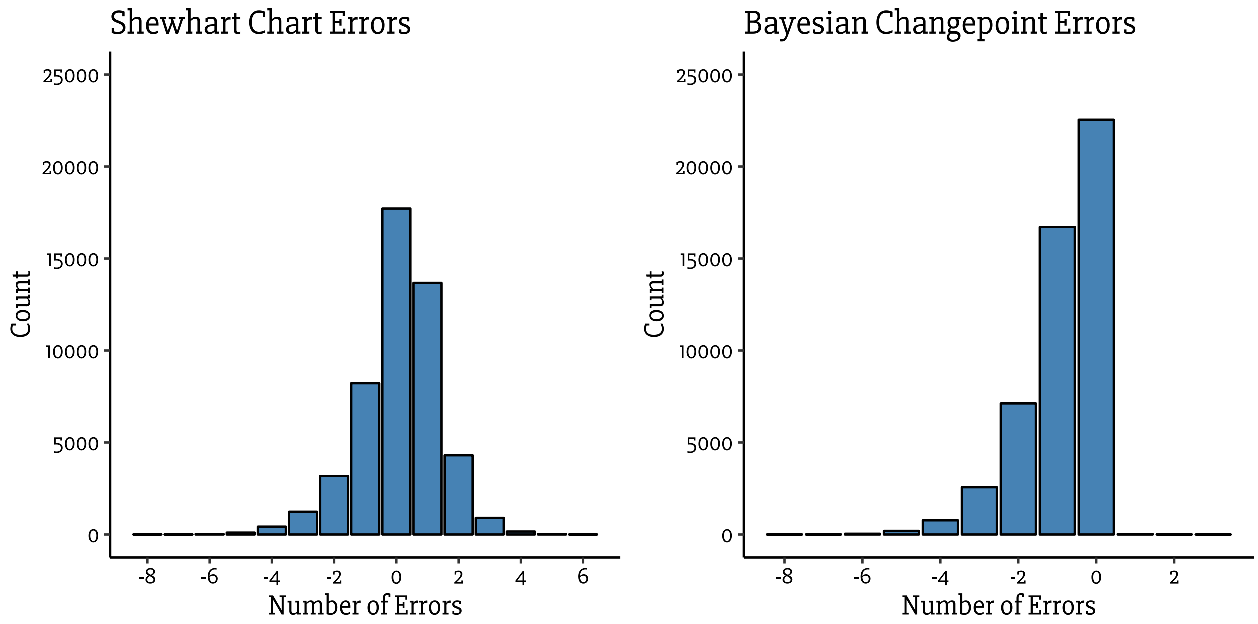

Figure 3: Overview of Monte Carlo simulation results. Errors correspond to the difference between the estimated number of coefficient changes in a single iteration and the estimated number of changes. For example, an error of -1 could occur if there are 3 actual coefficient changes, but only 2 are identified. On the other hand, an error of 2 could occur if there is 1 coefficient change, but 3 are identified. The error count represents the number of simulation iterations where each error count occurred

performance for all iterations at the allocated number of coefficient changes.

Table 1: **Mean-Squared Error by Number of Changes**

	Zero	One	Two	Three	Four	Five	Six +
Shewhart Chart	1.02	1.14	1.45	2.35	4.15	6.52	10.66
Bayesian Changepoint	0.004	0.53	1.75	3.70	6.49	10.04	14.39

We also include the mean-squared error at each number of changes in Table 1. As the time series becomes increasingly dynamic, the Shewhart chart's relative performance increases.[6] Due to its conservative retrospective nature Bayesian changepoint analysis outperforms our technique when the time series of coefficient estimates is stable. However, once we reach two changes, the Shewhart chart's mean-squared error and median errors are consistently lower. This trend only increases as we we introduce more coefficient changes. When there are four or

[6] The number of errors is counted as the number of detected changes minus the actual number of changes. Therefore a negative error count represents missed changes and a positive count represents mistaking noise for a change.

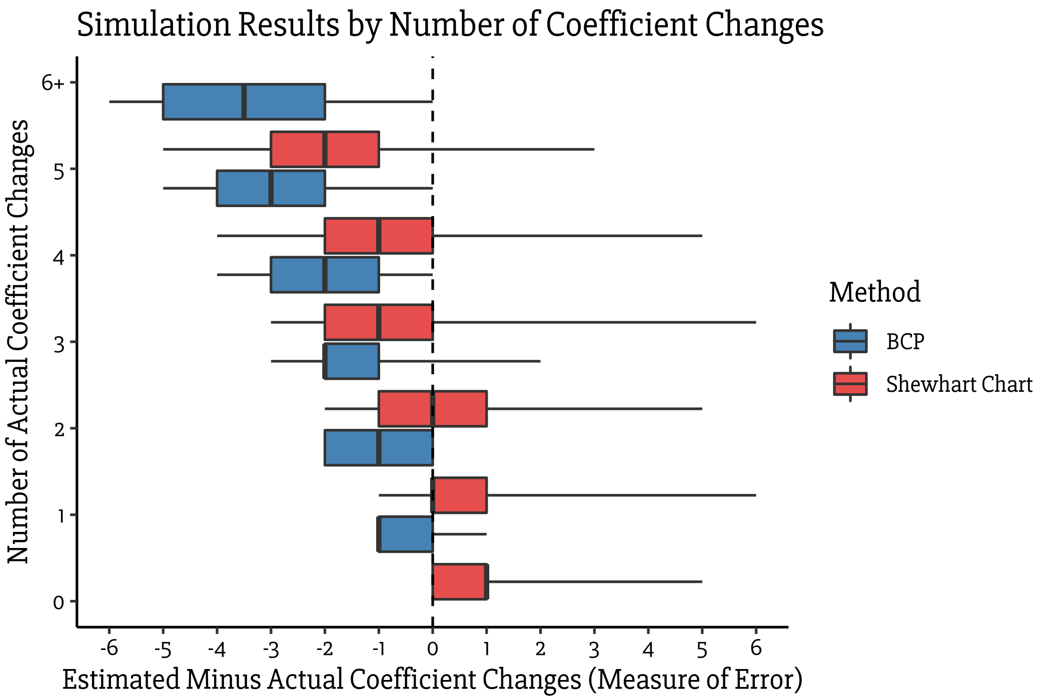

Figure 4: Simulation results for each method by the number of coefficient changes built into the data. Following Figure 3, The x-axis corresponds to the difference between the estimated and actual number of coefficient changes in a single iteration. The y-axis is the number of actual coefficient changes built into the simulated data in any iteration. This approach lets us compare the distribution of errors for each method, according to the number of coefficient changes in the simulated data. As the number of coefficient changes increases, the simulated data becomes more dynamic. The boxplots at each number of coefficient changes capture the summary statistics for each method's performance. The blue boxplots represent changepoint analysis and red represents the Shewhart chart.

more changes, the only summary statistic which bcp outperforms on is the maximum number of errors; yet even here the maximum number of errors for the Shewhart chart drops down to 3.

Last, we assess the two techniques by the absolute value of the distance between the detected and actual changes. For example, if a change occurred at time $t = 10$, but is estimated to have occurred at $t = 12$, then the distance between the estimated and actual change is 2. To do so, we run the Monte Carlo simulation with only one coefficient change occurring in each iteration, but at varying locations and at varying magnitudes. It is worth noting that this is an inherently uphill test for our technique, given bcp's strength with low-change time-series. However, assessing the distance between detected and actual changes is clear when there is only one change, as we can just take the absolute value of the difference between the closest detected change and the actual change. Turning to the results of this simulation, as table 8 demonstrates, *when it detects a change*, changepoint analysis is highly accurate (likely due to its retrospective, as opposed to prospective, nature), with a tendency to estimate the correct location. However, changepoint analysis failed to detect *over half* of the changes.[7] On the other hand, the Shewhart chart's median distance is only off by 1 time period.[8] In this simulation, therefore, both techniques encounter meaningful difficulties. Changepoint analysis is very accurate when it correctly estimates a change having occurred, but its conservative nature leads to failing to detect a change over 50% of the time. The Shewhart chart is much more effective at capturing that changes have occurred, but is less accurate with respect to a change's location. Future research may want to explore a means of either increasing the spatial accuracy of our prospective test or decreasing the tendency to miss changes in retrospective measures like bcp.

Table 2: **Distance Between Actual and Detected Change**

	Undetected	Min	First Quantile	Median	Mean	Third Quantile	Max
Shewhart Chart	721	0	0	1	9.77	9	98
Bayesian Changepoint	2605	0	0	0	0.11	0	83

Application

To demonstrate the model's applicability, we replicated Leeds' (2003) study on the relationship between alliance types and the initiation of Militarized Interstate Disputes (MIDs). Using data spanning from 1816 to 1944, Leeds fits a generalized estimating equation (Zorn, 2001) with terms for the association between various types of alliances – defensive alliances, offensive alliances, and neutrality pacts – and the initiation of a MID. Leeds argues that defensive alliances are negatively associated with MID initiation and that offensive alliances and neutrality pacts

[7] The simulation included 5000 iterations.

[8] The mean's distance is leveraged by the greater number of extreme errors.

are positively associated with MID initiation. In other words, defensive alliances deter conflict by raising the expected costs of war, offensive alliances encourage conflict by increasing a state's likelihood of winning a war, and neutrality pacts increase the likelihood of conflict by creating an expectation that potential adversaries will stay out of the war.

Though the paper is important and interesting, it fits a single model to over 100 years of data. Even if the general direction of the relationship between alliance types and conflicts holds constant over the entire time period, given the dynamic nature of the international system it is reasonable to expect that the magnitude of the relationship has varied meaningfully throughout history. In this regard, applying the Shewhart chart to the paper's models opens up the opportunity to ask when the relationship between alliances and conflict has changed. Taking this approach allows a researcher to ask 'under what conditions are alliances more versus less effective?' instead of solely focusing on whether alliances appear effective on average.

After replicating the paper's pooled model, we first employed our likelihood-ratio test to explore whether or not there is sufficient evidence to reject the hypothesis of effect stability. The test revealed sufficient evidence to reject the use of a single pooled model; the difference between the log-likelihood of the time-varying and pooled models was greater than that calculated in all bootstrapped samples. Second, in order to test for the number of changes and their location, we obtained coefficient estimates over time by repeatedly fitting a model with temporal subsets of the dataset. We produced estimates by: selecting twenty years of data[9], fitting the model, storing results, moving forward one year, fitting the model again, and so on.[10] This approach allowed us to obtain sufficient coefficient estimates for training the Shewhart chart while minimizing the likelihood of a parameter change taking place in training.[11] As Figure 5 shows, five coefficient estimates signaled a change according to the Shewhart chart. Each coefficient change is marked by a large red dot-whisker bar. These changes occurred during

[9] This length was chosen because we found it allowed for enough variation in the dependent variable to obtain reliable coefficient estimates, but maximized the number of coefficient estimates we could analyze.

[10] In other words, we fit a models from 1845 – 1865, 1846 – 1866, 1847 – 1867, ..., 1924 – 1944. We did not include estimates from 1816 – 1844 because these models produced unreliable estimates due to model separation.

[11] Note, including overlapping data windows is not a problem if the relationship is truly stable over time, which one assumes when fitting a single pooled model.

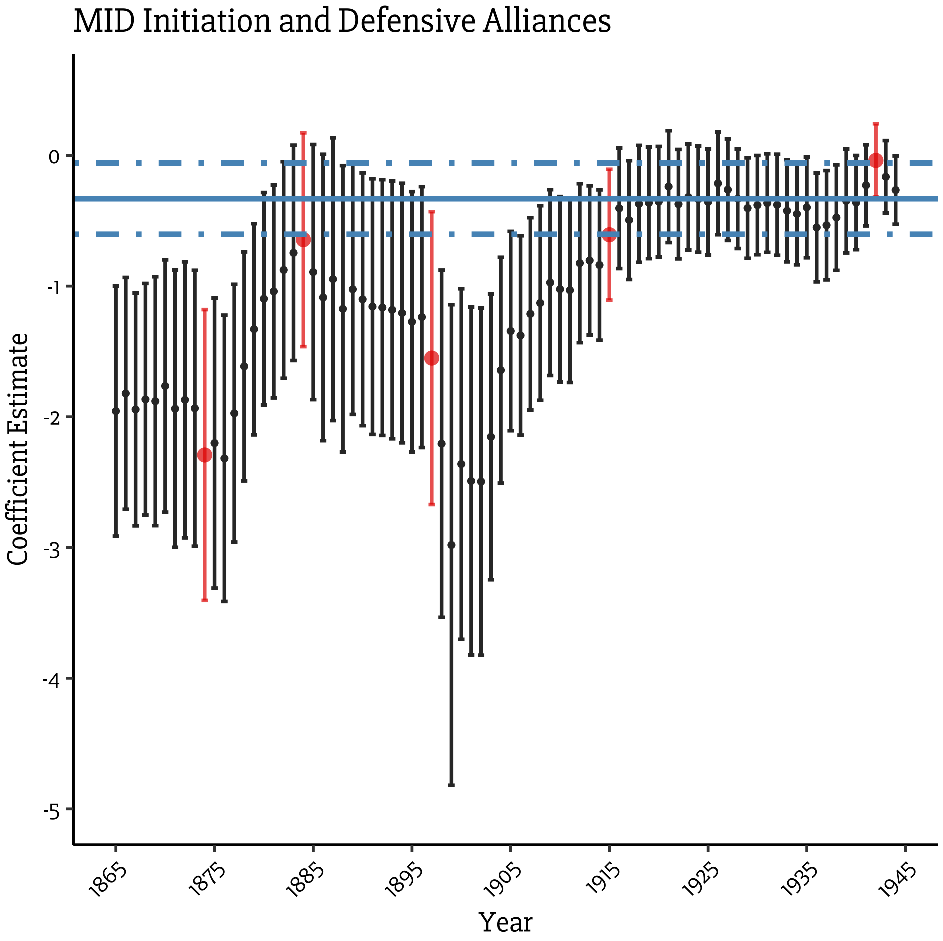

Figure 5: Replication and extension of Leeds (2003). Large red estimates correspond to a signaled change. The blue horizontal lines represent the pooled point estimate and its confidence interval. Note that the technique captures a change in the point-estimate, not the entire confidence interval. Accordingly, changes are detected even when confidence intervals overlap.

the following models: 1854 – 1874, 1864 – 1884, 1877 – 1897, 1895 – 1915, and 1922 – 1942. These estimates are promising in that they overlap well with reasonable historically-driven explanations. The first change corresponds with a relatively conflict-averse ear resulting from the Concert of Europe.[12] The second associates with various Prussian conflicts of German unification. The third corresponds with the aftermath of the wars of German unification. The fourth and fifth include the entirety of World War I and the United States' entry into World War II.

Moreover, the pooled estimate and its confidence interval are included in blue across the

[12] Note that this does not represent a conflict-free era; even the first model fit (1845 – 1865) includes the Crimean War.

figure. As the figure demonstrates, during the long-peace associated with the Concert of Europe defensive alliances were associated with far greater deterrent effects than a pooled model suggests. This relationship dissipates, if not disappears with WWI, the interwar period, and WWII.[13] Interestingly, this suggests that the pooled estimate is generally somewhat conservative and that data around WWII exhibits greater leverage than expected.

Discussion and Conclusion

We have claimed and sought to demonstrate that (a) effect dynamicism is common in temporal political science data, (b) that such dynamicism can lead to faulty inference when it exists but goes unmodeled, (c) that standard regression-type models do not accommodate dynamic effects when they produce single coefficients for a given period of observation, and that (d) our SPM approach makes detecting and exploring effect dynamicism easy. The ease by which we can now explore effect dynamicism opens new theoretical horizons for theory and our understanding of important processes in political science. Our results also carry the troubling implication that a significant proportion of established results in our field may be incorrect or partially correct due to effect dynamicism.

Because the technology for fast, easy, and stable exploration of temporal dynamics was not available until now, one can hardly blame researchers for assuming temporal stability; it was an assumption forced upon them by limitations in our methodological toolkit. Now that such dynamic analysis is simple, we believe that the assumption of stable effects should not be the norm, but rather an exception proved through empirical investigation. Our technique presents an easy way to do so and to test evidence regarding effect stability in a familiar framework. Perhaps most importantly, considering effect dynamicism opens up a new dimension in which strong theory may continue to advance our state of knowledge.

[13] Of course, applying the technique appropriately assumes that one is receiving accurate parameter estimates – that the underlying statistical model is fit appropriately. There are reasons to be concerned about a dyadic research design in this context, particularly given the amount of evidence that exists for network effects in alliances. (Cranmer and Desmarais, 2016, e.g.)

We introduced a statistical monitoring strategy for detecting and diagnosing effect heterogeneity. Our proposed method answers two important questions about heterogeneity in generalized linear models. First, we utilize a bootstrap procedure for identifying temporal heterogeneity to formally test whether a pooled model is appropriate for the panel data being analyzed. Next, we monitor the effect coefficients describing the relationship between the predictors on the response individually to determine exactly when and for which predictors heterogeneity is present. We demonstrate the utility of our methods on a widely cited article on the relationship between alliances and the initiation of militarized interstate disputes. Though this article's key relationships do not change in direction over time, they vary in magnitude and change at important inflection points. Indeed, the pooled model *understates* the impact of alliances at various points in history, demonstrating one of the risks researchers face when pooling effects, as opposed to modeling them dynamically over time.

By implementing our SPM technique in free and easy-to-use companion software, we endeavor to make checking for effect dynamicism as close to trivial for the applied researcher as possible. Diligent checking for dynamic effects not only helps the researcher avoid inferential errors, but is useful for answering substantively interesting questions and testing temporally nuanced theories.

References

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

N. Beck. Time-series–cross-section data: What have we learned in the past few years? *Annual review of political science*, 4(1):271–293, 2001.

N. Beck. Time-series-cross-section methods. *Oxford Handbook of Political Methodology*, pages 475–93, 2008.

N. Beck and J. N. Katz. What to do (and not to do) with time-series cross-section data. *American political science review*, 89(3):634–647, 1995.

N. Beck and J. N. Katz. Nuisance vs. substance: Specifying and estimating time-series-cross-section models. *Political analysis*, 6:1–36, 1996.

N. Beck, J. N. Katz, and R. Tucker. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, pages 1260–1288, 1998.

A. Bell and K. Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153, 2015.

D. S. Bennett and A. C. Stam. *The behavioral origins of war*. University of Michigan Press, 2009.

J. Benneyan, R. Lloyd, and P. Plsek. Statistical process control as a tool for research and healthcare improvement. *Quality and Safety in Health Care*, 12(6):458–464, 2003.

J. M. Box-Steffensmeier and B. S. Jones. *Event history modeling: A guide for social scientists*. Cambridge University Press, 2004.

J. M. Box-Steffensmeier, J. R. Freeman, M. P. Hitt, and J. C. Pevehouse. *Time series analysis for the social sciences*. Cambridge University Press, 2014.

P. T. Brandt and T. Sandler. Hostage taking: Understanding terrorism event dynamics. *Journal of Policy Modeling*, 31(5):758–778, 2009.

B. F. Braumoeller. *The great powers and the international system: systemic theory in empirical perspective*. Cambridge University Press, 2013.

D. B. Carter and C. S. Signorino. Back to the future: Modeling time dependence in binary data. *Political Analysis*, 18(3):271–292, 2010.

T. S. Clark and D. A. Linzer. Should I use fixed or random effects? *Political Science Research and Methods*, 3(2):399–408, 2015.

S. J. Cranmer and B. A. Desmarais. A critique of dyadic design. *International Studies Quarterly*, 60(2):355–362, 2016.

S. J. Cranmer, T. Heinrich, and B. A. Desmarais. Reciprocity and the structural determinants of the international sanctions network. *Social Networks*, 36:5–22, 2014.

S. De Boef and L. Keele. Taking time seriously. *American Journal of Political Science*, 52(1): 184–200, 2008.

J. De Brabanter, K. Pelckmans, J. Suykens, and B. De Moor. Generalized likelihood ratio statistics based on bootstrap techniques for autoregressive models. *IFAC Proceedings Volumes*, 39 (1):790–795, 2006.

W. Enders. Applied econometric time series, by walter. *Technometrics*, 46(2):264, 2004.

C. Erdman, J. W. Emerson, et al. bcp: an r package for performing a bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

J. Fan and W. Zhang. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731, 2000.

J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008.

J. Fan, C. Zhang, and J. Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Annals of Statistics*, pages 153–193, 2001.

R. J. Franzese Jr and J. C. Hays. Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis*, 15(2):140–164, 2007.

J. R. Freeman, J. C. Hays, and H. Stix. Democracy and markets: The case of exchange rates. *American Journal of Political Science*, pages 449–468, 2000.

A. Gelman and J. Hill. *Data analysis using regression and multilevel hierarchical models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

J. Gill. *Bayesian methods: A social and behavioral sciences approach*, volume 20. CRC press, 2014.

K. S. Gleditsch and M. D. Ward. War and peace in space and time: The role of democratization. *International Studies Quarterly*, 44(1):1–29, 2000.

J. Golub. Survival analysis. In *The Oxford handbook of political methodology*. Oxford University Press, 2008.

D. P. Green, S. Y. Kim, and D. H. Yoon. Dirty pool. *International Organization*, 55(2):441–468, 2001.

T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

D. R. Hoover, J. A. Rice, C. O. Wu, and L.-P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

R. R. Huckfeldt, C. W. Kohfeld, and T. W. Likens. *Dynamic modeling: An introduction*. Number 27 in Quantitative Modeling in the Social Sciences. Sage, 1982.

L. Jenke and C. Gelpi. Theme and variations: Historical contingencies in the causal model of interstate conflict. *Journal of Conflict Resolution*, pages 1–23, 2016.

L. Keele and N. J. Kelly. Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political analysis*, 14(2):186–205, 2005.

G. King. *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press, 1998.

J. Lee, G. Li, and J. D. Wilson. Varying-coefficient models for dynamic networks. *arXiv preprint arXiv:1702.03632*, 2017.

B. A. Leeds. Do alliances deter aggression? the influence of military alliances on the initiation of militarized interstate disputes. *American Journal of Political Science*, 47(3):427–439, 2003.

J. F. MacGregor and T. Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3):403–414, 1995.

R. McCleary, R. A. Hay, E. E. Meidinger, and D. McDowall. *Applied time series analysis for the social sciences*. JSTOR, 1980.

S. M. Mitchell, S. Gates, and H. Hegre. Evolution in democracy-war dynamics. *Journal of Conflict Resolution*, 43(6):771–792, 1999.

D. C. Montgomery. *Introduction to statistical quality control*. John Wiley and Sons, Inc, 7 edition, 2013.

D. C. Montgomery and J. B. Keats. *Statistical Process Control in Manufacturing*. Marcel Dekker, 1991.

M. D. Nieman. Moments in time: Temporal patterns in the effect of democracy and trade on conflict. *Conflict management and peace science*, 33(3):273–293, 2016.

J. S. Oakland. *Statistical process control*. Routledge, 2007.

J. H. Park. Analyzing preference changes using hidden markov item response theory models. *Handbook of Markov chain Monte Carlo: Methods and applications, eds. Galin Jones, Steve Brooks, Andrew Gelman, and Xiao-Li Meng. Boca Raton, FL: Chapman and Hall/CRC*, 2011a.

J. H. Park. Changepoint analysis of binary and ordinal probit models: An application to bank rate policy under the interwar gold standard. *Political Analysis*, 19(2):188–204, 2011b.

J. H. Park. A unified method for dynamic and cross-sectional heterogeneity: Introducing hidden markov panel models. *American Journal of Political Science*, 56(4):1040–1054, 2012.

T. Plümper and V. E. Troeger. Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis*, 15(2):124–139, 2007.

K. Scheve and D. Stasavage. Institutions, partisanship, and inequality in the long run. *World Politics*, 61(2):215–253, 2009.

A. Spirling. “Turning points” in the Iraq conflict: Reversible jump markov chain monte carlo in political science. *The American Statistician*, 61(4):315–320, 2007.

G. J. Wawro and I. Katznelson. Designing historical social scientific inquiry: how parameter heterogeneity can bridge the methodological divide between quantitative and qualitative approaches. *American Journal of Political Science*, 58(2):526–546, 2014.

J. D. Wilson, N. T. Stevens, and W. H. Woodall. Modeling and estimating change in temporal networks via a dynamic degree corrected stochastic block model. *arXiv preprint arXiv:1605.04049*, 2016.

W. H. Woodall and D. C. Montgomery. Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31(4):376–386, 1999.

W. H. Woodall and D. C. Montgomery. Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, 46(1):78–94, 2014.

W. H. Woodall, M. J. Zhao, K. Paynabar, R. Sparks, and J. D. Wilson. An overview and perspective on social network monitoring. *IISE Transactions*, 49(3):354–365, 2017.

C. J. Zorn. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, pages 470–490, 2001.