PA

A Permutation-Based Changepoint Technique for Monitoring Effect Sizes

Daniel Kent[1], **James D. Wilson**[2] **and Skyler J. Cranmer**[3]

[1] *Department of Political Science, The Ohio State University, 2140 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, USA.*
Email: kent.249@osu.edu
[2] *Department of Mathematics and Statistics, University of San Francisco, San Francisco, CA 94117, USA.*
Email: jdwilson4@usfca.edu
[3] *Department of Political Science, The Ohio State University, 2140 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, USA.*
Email: cranmer.12@osu.edu

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 analyses of time-varying processes, we develop a two-stage procedure based upon techniques for permutation testing and statistical process monitoring. Given time series cross-sectional data, we break the role of time through permutation inference and produce a null distribution that reflects a time-invariant data generating process. The null distribution then serves as a stable reference point, enabling the detection of effect changepoints. In Monte Carlo simulations, our randomization technique outperforms alternatives for changepoint analysis. A particular benefit of our method is that, by establishing the bounds for time-invariant effects before interacting with actual estimates, it is able to differentiate stochastic fluctuations from genuine changes. We demonstrate the method's utility by applying it to a popular study on the relationship between alliances 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 questions about such changes.

Keywords: changepoint analysis, permutation tests, time series cross-sectional data

1 Introduction

Many important social and political processes—such as international conflict, democratization, voting behavior, and public policy implementation—manifest over time. Since these are inherently longitudinal processes, it is natural that their temporal dynamics be a major focus for researchers. In particular, when working with time series cross-sectional (TSCS) data, *relationships* between variables are likely to vary over time. Such time-varying relationships stand in stark contrast to the implicit assumption of stable effects built into models that estimate a single set of coefficients for data spanning a wide range of time. Indeed, correctly detecting where relationships between variables change is not only important for making accurate inferences, it also provides an opportunity to explore substantively interesting questions about why relationships change at some points in time but not others. In order to do so, we propose a novel technique for estimating the probability that a coefficient's magnitude changes at any point in time when fitting generalized linear models with TSCS data.

We approach time-varying relationships as a changepoint problem (e.g., Barry and Hartigan 1993; Erdman and Emerson 2007; Killick and Eckley 2014; Blackwell 2018), where the object of interest is a time series of coefficients with a mean that changes an unknown number of times at unknown locations. However, unlike changepoint approaches that seek to find the set of partitions within a time series that minimize a cost function, we propose a novel two-step technique that uses a permutation test to inform a statistical process monitoring (SPM) procedure.[1] More specifically,

Political Analysis (2020)

DOI: 10.1017/pan.2020.44

Corresponding author
Daniel Kent

Edited by
Sunshine Hillygus

1 For similar applications that also overview the SPM literature, see Ge, Song, and Gao (2013), Montgomery and Keats (1991), Montgomery (2013), Wilson, Stevens, and Woodall (2019), and Woodall *et al.* (2017).

we first approximate a null distribution of time-invariant coefficients through permutation inference. The resulting distribution sets the size of a moving window that scans the time series and returns the probability of a changepoint for each coefficient.

Notably, our discussion is not the first to highlight the importance of accurately modeling time-varying coefficients.[2] Nor are we alone in discussing the inferential challenges associated with dynamic data.[3] But we find in Monte Carlo simulations that our technique provides substantial advantages over other comparable changepoint methods. The method's accuracy derives from its use of data-driven bounds for stable behavior, which minimize the risk of overfitting when detecting changepoints.

After presenting the method in detail and discussing Monte Carlo simulation results, we apply the technique to a popular study of the relationship between alliances and militarized interstate dispute initiation. We find that the study's general claims about deterrence hold, but that the relationship's magnitude changes at multiple historically interesting times. These changepoints provide an example of how the method can raise conceptual questions about why a relationship varies when it does. We close with some notes of caution and recommendations for best-practices.

2 Monitoring Generalized Linear Models

In TSCS data, one observes N units repeatedly over T time periods. Data can also be represented by the sequence $\{(X_{it}, Y_{it}),\ i = 1, \ldots, N,\ t = 1, \ldots, T\}$, where $Y_{it} \in \mathbb{R}^n$ is the outcome response variable for observation i at time t, and X_{it} is an $n \times p$ matrix of p predictors at time t. In many cases, the goal is to estimate the relationship between Y_{it} and X_{it}. For a *varying-coefficient* generalized linear model (Hastie and Tibshirani 1993), one does this by treating Y_{it} as a random variable whose mean μ_{it} relates to the predictors X_{it} as follows:

$$g(\mu_{it}) = X_{it}\boldsymbol{\beta}_t, \quad i = 1, \ldots, N \quad t = 1, \ldots, T. \tag{1}$$

where a separate vector of p coefficients, $\beta_t = (\beta_{1t}, \beta_{2t}, ..., \beta_{pt})$, is fit for each time period and $g(\cdot)$ is an appropriately chosen link function that maps μ_{it} to a linear function of the predictors X_{it}. Common choices of $g(\cdot)$ include the identity, logit, and natural log function, corresponding to multivariate linear, logistic, and Poisson regression, respectively.[4]

In the case of a *pooled* glm, one is quantifying the *average* effect of X_{it} on the mean of Y_{it}, requiring a static relationship between X_{it} and Y_{it} through time. The average relationship is accurate across each time period and equivalent to a time-varying model if $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_T$. For many applications this assumption simply does not hold, as the relationship between X_{it} and Y_{it} is dynamic for at least one key variable.[5] In light of such concerns over detecting where relationships change and understanding why, we provide a technique that answers two questions about the relationship between two variables:

1. Is the effect β_t constant through time?
2. If β_t is not constant across all time points, then when does the effect significantly differ from previous periods?

2 Examples include, but are not limited to: Braumoeller (2013), Cranmer, Heinrich, and Desmarais (2014), Jenke and Gelpi (2016), Thurner *et al.* (2018), and Wawro and Katznelson (2014).

3 E.g., Beck (1983), Beck, Katz, and Tucker (1998), Beck (2001, 2008), Box-Steffensmeier and Jones (2004), Box-Steffensmeier *et al.* (2014), Carter and Signorino (2010), De Boef and Keele (2008), Gelman and Hill (2007), Gill (2014), Golub (2008), Huckfeldt, Kohfeld, and Likens (1982), Jenke and Gelpi (2016), King (1998), Mitchell, Gates, and Hegre (1999), Nieman (2016), Park (2012), Wawro and Katznelson (2014), and Zorn (2001).

4 Note that the formulation in Equation (1) is amenable to more complicated extensions, such as fixed or random effects, lagged dependent variables, first differences, various standard errors, and more. But if working with TSCS data, there always is a concern that, however extensive one's model, if a single coefficient is fit for multiple time periods, then that estimate may miss important variation over time.

5 This issue is first raised in Political Science by Beck (1983).

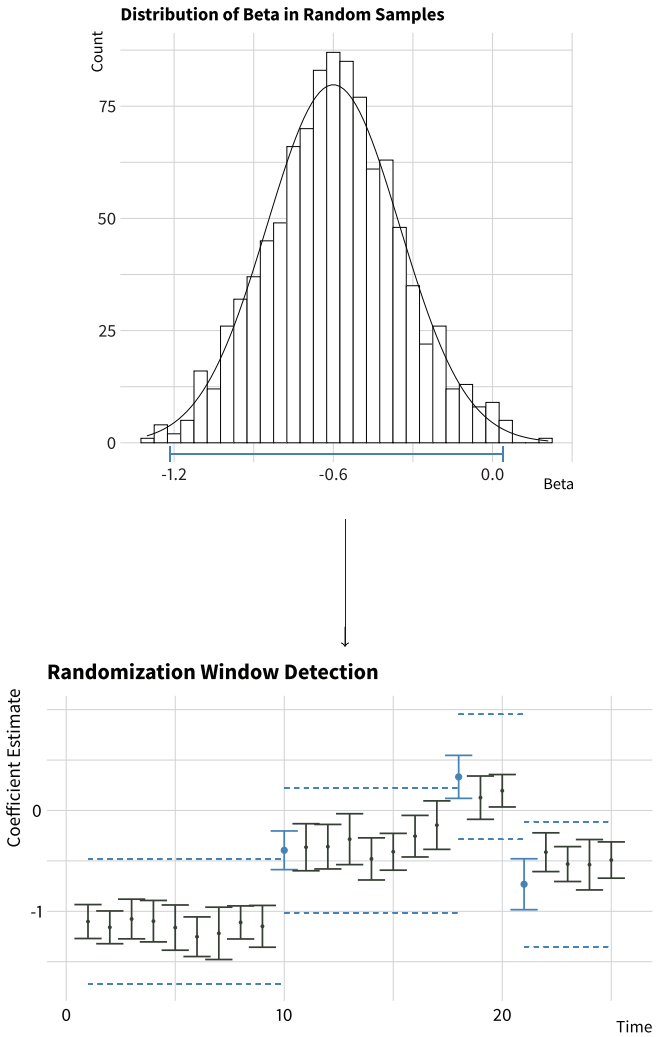

Figure 1. Display of constructing and applying the moving window. The top distribution includes the estimated values of $\hat{\beta}$ from random temporal samples. The blue bar at the base of the distribution is the estimated tolerance region, $\mathcal{R}$. The lower figure then is the time series of coefficient estimates for the simulated data. The tolerance region begins at time = 1 and once a coefficient's point-estimate falls outside of the region a change is signaled. That change is colored blue. A new region is built around the change with the same size but a new center.

2.1 Detecting Change Locations

Consider a time series of coefficient estimates for a generalized linear model in Equation (1):

$(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_T)$. If the relationship is constant across any period of time, then each observed $\hat{\beta}_t$ estimate within that range will vary around a stable mean value. However, if the relationship changes, then the subsequent observed $\hat{\beta}_t$ estimates will vary around a new value. For any two variables, a visual demonstration of such a time-varying relationship can be found at the bottom of Figure 1. Detecting changepoints in such a time series is a statistical challenge because predictions are unverifiable and all time points are candidates for change, meaning there are an unknown number of partitions in the time series with unknown locations. In this sense, one is faced with an *unsupervised* exercise of estimating $\Pr(\text{Change}_t = 1)$ for each point in time.

In order to detect where changes have occurred, we draw on the SPM literature and apply a moving window to the time series of coefficients $\hat{\beta}_t$. SPM approaches take a statistic S_t (in our application coefficient estimates), or more generally a vector of statistics $\boldsymbol{S}_t$, that provides a summary of a dataset $\{(X_t, Y_t),\ t = 1, \ldots, T\}$. Perhaps the most common SPM approach for identifying multiple changes is a moving-window control chart (Ge *et al.* 2013). In a typical moving-

window approach, one produces a *tolerance region*, $\mathcal{R}$, based upon the values of S_t in a training period. For each new set of observations (X_t, Y_t), S_t is calculated. The new data is deemed "typical" if $S_t \in \mathcal{R}$ and deemed "anomalous" otherwise. Once a data point is labelled anomalous, a new tolerance region is produced.

Our approach determines the size of $\mathcal{R}$ through permutation inference before monitoring the time series. By repeatedly sampling a random set of time periods without replacement, extracting all data in those time points, and estimating the coefficient of interest, we are able to approximate a null distribution of time-invariant coefficients. Because the time periods are sampled at random, any observed differences across estimates are not expected to be due to the time periods selected. The permutation-based null distribution informs the tolerance region's size by estimating how much coefficient variation occurs when the relationship is time-invariant. More formally, this permutation procedure is executed in the following steps.

Let $Y_{it} = g^{-1}(X_{it}\beta_t) + \varepsilon_{it}$, where $i = 1, ..., N$ and $t = 1, ..., T$. Moreover, let L represent the number of sampling iterations and J be the size of each random sample, where a sample is a set of time periods. The following four steps are repeatedly carried out:

Constructing a Null Distribution for $\widehat{\beta}$

- For each $\ell \in \{1, ..., L\}$:
  - Randomly sample J time periods without replacement, so that each $j \in \{1, ..., T\}$
  - For each $j \in \{1, ..., J\}$ extract (X_j, Y_j)
  - Estimate $\hat{\beta}$, where $Y_{ij} = g^{-1}(X_{ij}\beta) + \varepsilon_{ij}$
  - Store $\hat{\beta}^{(\ell)} = \hat{\beta}$

The sampled values $\{\hat{\beta}^{(\ell)} : \ell = 1, \ldots, L\}$ serve as a reference distribution against which we test for change. Specifically, we construct the tolerance region $\mathcal{R}$ using a *Shewhart control chart*, where the tolerance region for element β_{kt} at time t is $\hat{\beta}_{kt} \pm 3\sigma_k$, $k = 1, ..., p$ where σ_k is the standard deviation of the estimated coefficients $\{\hat{\beta}_k^{(1)}, \ldots, \hat{\beta}_k^{(L)}\}$, and t is the first time point in the tolerance region. We note that the Shewhart chart is one of many possible choices of a control chart from the SPM literature. 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).[6] We subsequently refer to our technique as a Shewhart chart.

For parsimony, assume that the subsequent monitoring procedure is carried out on a single variable, whose estimated coefficient at time t is $\hat{\beta}_t$.[7] Next, a prespecified probability cutoff, ρ (0.5 by default), is selected to represent the necessary probability for $\hat{\beta}_t$ to signal a change. The probability that $\hat{\beta}_t$ signals a change is then the percent of the estimated sampling distribution that is greater than or less than $\mathcal{R}$, depending on whether the point estimate is above or below the mean value within $\mathcal{R}$. In other words, the probability of change for each time point, $\Pr(\text{Change}_t = 1)$, is the density of the estimated sampling distribution for $\hat{\beta}_t$ outside of $\mathcal{R}$. Though easily changed, as a default if a coefficient's point estimate is not inside $\mathcal{R}$, then we reflect most SPM approaches and label an estimate as a change because more than half of the sampling distribution is outside $\mathcal{R}$. This follows standard practice with logistic regression, where a predicted probability greater than

6 Our strategy of monitoring the coefficients of a fitted model is related to the work of Wilson *et al.* (2019), 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. The primary motivation for using $\pm 3\sigma$ stems from prediction intervals being larger than confidence intervals. In addition, rather than $\pm 2\sigma$, we opt for $\pm 3\sigma$ because of the risk of multiple comparisons; repeatedly comparing a threshold to observations risks mistaking noise for change, so we favor a larger window.

7 This stands in contrast to the discussion thus far which has been about a general discussion of monitoring a vector of p coefficients over time.

0.50 is generally classified as 1, rather than 0, because 1 is more likely than not. If the estimated $\Pr(\text{Change}_t = 1) > \rho$, then $\hat{\beta}_t$ is labelled a changepoint. Formally, this process is carried out as:

Monitoring a Coefficient, $\hat{\beta}_t$

- For each $t \in \{1, \ldots, T\}$, with a prespecified threshold ρ:
  - If $t = 1$:
    - Let $\mu_k = \hat{\beta}_t$, where $k \in \{1, \ldots, K\}$ and K is the final number of partitions.
    - Let $\mathcal{R}_k = \mu_k \pm 3\sigma$, where $\mathcal{R}_k$ is the tolerance region for partition k and σ is the standard deviation of the estimated null distribution of time-invariant effects.
  - Else:
    - If: $\Pr(\hat{\beta}_t \in \mathcal{R}) \leq \rho$, where $\Pr(\hat{\beta}_t \in \mathcal{R})$ is the percent of the sampling distribution for $\hat{\beta}_t$ within $\mathcal{R}$:
      - $\hat{\beta}_t$ is *not* labelled a changepoint and $\Pr(\hat{\beta}_t \in \mathcal{R})$ is stored.
    - Else:
      - $\hat{\beta}_t$ is labelled a changepoint and $\Pr(\hat{\beta}_t \in \mathcal{R})$ is stored.
      - A new tolerance region, $\mathcal{R}_k$ is calculated, where $\mathcal{R}_k = \hat{\beta}_t \pm 3\sigma$.

2.2 Minimizing Overfitting

Given the range of options for changepoint analysis (Aminikhanghahi and Cook 2017), why is the proposed technique favorable? The method's conceptual strength stems from its ability to avoid overfitting. As discussed in Haynes, Eckley, and Fearnhead (2017) and Killick and Eckley (2014), changepoint analyses are easily susceptible to mistaking noisy fluctuations for genuine change, inspiring a literature on penalties meant to offset this risk.[8] For example, in a popular R package, estimates are produced by minimizing the following: [9]

$$\sum_{i=1}^{m+1} \left[C(y_{(\tau_{i-1}+1):\tau_i}) \right] + \beta f(m). \tag{2}$$

Here m is the number of proposed partitions, y is the ordered data sequence, τ_{i-1} and τ_i are the start and end of candidate partitions, C is a cost function, and $\beta f(m)$ is a penalty that guards against overfitting. For example, the penalty value we use in simulation with this software is $2 \times \log(n)$, which is equivalent to the BIC/SIC penalty. In general, $\beta f(m)$ is a tunable parameter which dictates the number of changepoints identified.

The emphasis placed on penalty inclusion in popular research and software captures the challenge posed by overfitting. However, in these applications, because true changepoints are unknown one cannot evaluate the quality of their exogenously set penalty. This leaves the researcher without knowledge of whether or not they are applying a penalty that is too strict or too light. We sidestep this tuning problem altogether through the permutation-based null distribution, which lets the data inform a threshold for differentiating noise from actual change.

3 Simulation Results

In order to test our method's performance we conduct a Monte Carlo simulation, evaluating the Shewhart chart's performance at each iteration and comparing it to alternative methods for changepoint analysis.[10] When evaluating the Shewhart chart across iterations, we compare it to

8 The approach is similar to a LASSO regression, where a regularization parameter, λ, imposes a penalty that shrinks coefficients toward zero, emphasizing strong relationships that are resilient to the penalty's imposition.
9 One can find a more detailed walkthrough in pages 2–5 of Killick and Eckley (2014).
10 Replication data and code can be found in Kent, Wilson, and Cranmer (2020a) and Kent *et al.* (2020b).

Table 1. Root mean squared error by number of changes.

	Zero	One	Two	Three	Four	Five	Six
Shewhart chart	1.28	0.71	1.12	1.86	2.68	3.49	4.32
Bayesian changepoint	1.25	1.18	1.67	2.27	3.17	3.86	4.85
Changepoint	0.06	0.81	1.56	2.38	3.31	4.14	5.12
CUSUM	0.00	1.00	2.00	3.00	4.00	4.99	6.00

BCP analysis (Erdman and Emerson 2007), the non-BCP techniques referenced in Equation (2), and a cumulative sum control chart (CUSUM) test. These alternative methods and accompanying R packages are chosen because they cover most approaches to changepoint analysis, have reasonable runtimes, and do not require that the user prespecifies the number of changepoints.[11]

The steps of each simulation iteration were specified so that data for a varying-coefficient linear model, $Y_{it} = X_{it}\beta_t + \varepsilon_{it}$, was generated with a prespecified number of changes in β_t for one predictor, x_{1t}, with all other relationships constant across the data. For each iteration a regression for every time period was fit, all methods were applied to estimates for the time-varying coefficient, and the difference between the true and estimated number of changes was stored. Across iterations, σ (the standard deviation for the distribution which coefficients were drawn from) was varied in magnitude when the time-varying coefficient β_t was generated, manipulating noise within each stable portion of time series. Specifically:

Monte Carlo Simulation Steps

- For each iteration:
  - The number of changepoints, P, is drawn from a truncated normal distribution.[12]
  - The location(s) of parameter changes are determined through P draws from a discrete uniform distribution, $U(1, T)$.
  - The mean value for each partition, μ_p, is drawn from a standard normal distribution, for $p = \{1, \ldots, P+1\}$.
  - Data for each time period, t is generated through a linear model where each coefficient value for the time-varying coefficient β_t is drawn from a normal distribution with mean μ_p and standard deviation σ.
  - All changepoint techniques are applied to the time series of coefficient estimates $\hat{\beta}_t$ for the time-varying relationship.
  - The difference between the actual and estimated number of changes is recorded.

Table 1 provides a broad summary of the simulation's results, where iterations are compared by the true number of coefficient changes. For each iteration, a method's error is calculated as the difference between the estimated and true number of changes, meaning errors closer to 0 represent better performance. For example, if there are two changes but only one is detected, then the error would be −1. Across each number of changes, with the time series growing increasingly dynamic as the number increases, the Shewhart chart has the lowest mean squared error, except for when there are zero changes.[13] Though, and we return to this before concluding, the more dynamic the time-series, then, regardless of method chosen, the more difficult accurate estimation becomes.

11 This is an unfortunate feature of MCMCpack's changepoint functionality, which requires specifying the number of changepoints in advance and then returns the most likely locations.

12 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.

13 On cases with zero changes, the technique uses 3σ from the distribution of time-invariant effects to inform the Shewart chart's bounds, or 99% of the estimated variation. Mistaking 1 in 50, or 98%, of the stable cases for a change follows expectations, providing additional evidence that the method performs as expected.

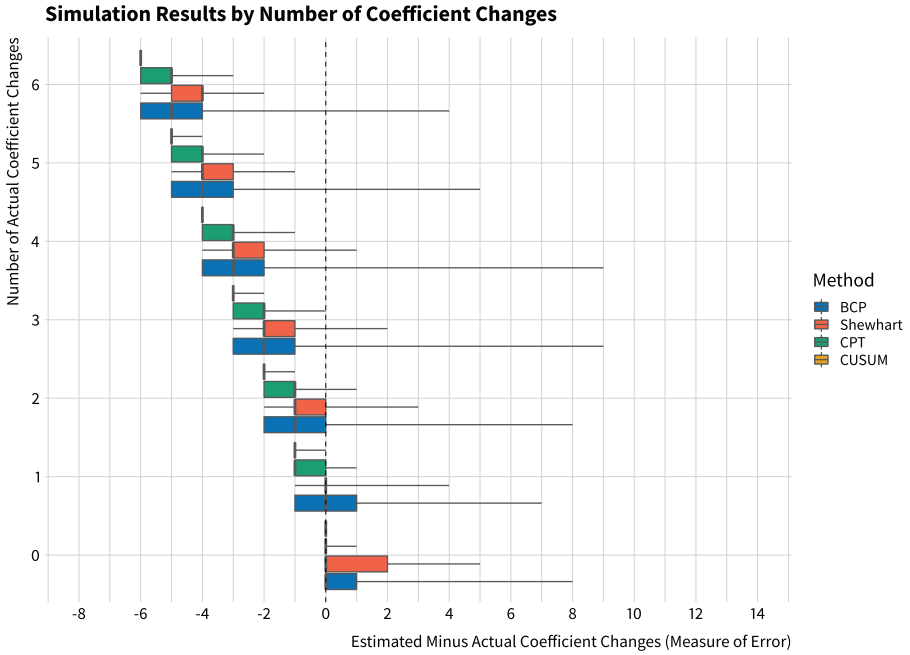

Figure 2. Simulation results for each method stratified by the true number of coefficient changes. The *x*-axis corresponds to the difference between the estimated and actual number of coefficient changes. The y-axis is the true number of coefficient changes in the simulated data. The boxplots capture the summary statistics for each method's performance. Results from BCP analysis are in blue, the Shewhart chart's results are in red, changepoint is in green, and CUSUM in gold. Whiskers are generally one-sided because each method only tends to have especially large errors when they overfit and mistake noise for changepoints, leading to a large positive difference between the estimated and true number of changes.

Moving to a visual representation, Figure 2 displays a summary of each method's performance across the simulation. Bayesian changepoint (BCP) analysis, which does not include a penalty, has the greatest tendency to overfit, with cases where the maximum number of errors is systematically far greater than 0. In cases where the error term is positive, the method of interest returns multiple false positives—detecting changepoints that do not exist. Considering the other methods, CUSUM tests are overly conservative and as the number of changepoints increases, the more they tend to treat the time series as having a stable mean but high variance. Changepoint analyses with an imposed penalty term avoid overfitting, having few instances of positive errors. But, the penalty appears to impose systematically too strict of a standard, rendering errors further from 0 in all cases except for those without any changepoints. Conversely, the Shewhart chart's performance is closest to 0 across most conditions.

Last, we assess the average run length—the average time it takes for a change to be detected, *once it has occurred*. To estimate this, we ran a Monte Carlo simulation where only one change occurs, but the magnitude of this change and its location varies. We then applied all techniques to the resulting time series of coefficient estimates, recording whether a change was detected and, if detected, the time elapsed after the change until detection. We present the results of the simulation in Figure 3 and Table 2. CUSUM and non-BCP tests are left out of these descriptive statistics because they tend to fail to detect a change in these situations. Visually, the results look similar across techniques. A closer look at the summary statistics reveals a more nuanced story, however. While the two methods perform similarly *when a change is detected*, with a lower mean for the Shewhart Chart but an equivalent median for both, the false-negative rate (undetected changes) is far higher for BCP analysis. Because there are so many more undetected changes for the Bayesian approach, its bars in Figure 3 are shorter than those for the Shewhart chart.

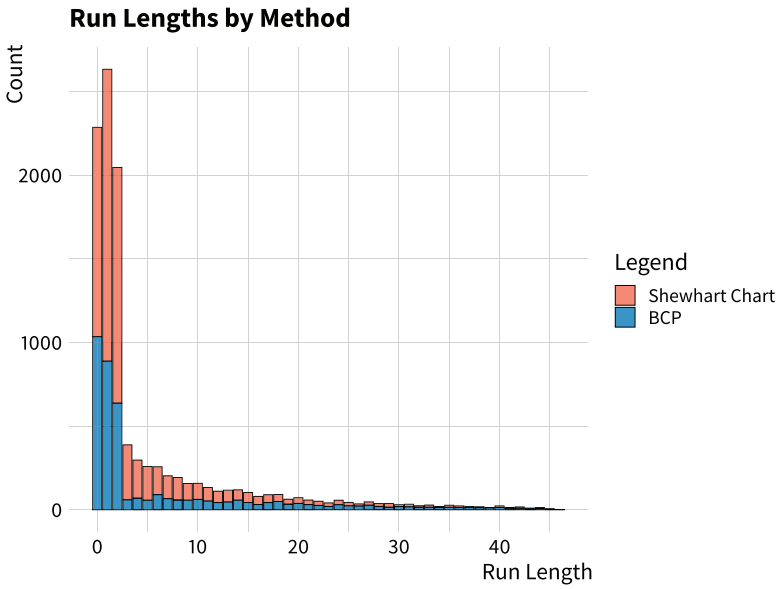

Figure 3. Distribution of run lengths for each method in a Monte Carlo simulation where only one change occurs. CUSUM and non-BCP results are not included because the methods tend to be too conservative and miss the change altogether. Results from the Shewhart chart are in red and BCP in blue. Run length is the amount of time between when a change occurs and the change is detected. 10,000 iterations were run in total. Cases where no change is detected are not included but noted in Table 2.

Table 2. Summary statistics for simulations with one change. The statistics represent: whether a change was detected and, if detected, the amount of time after a change before it is detected.

	Min	First quantile	Median	Mean	Third quantile	Max	Undetected
Shewhart chart	0	1	2	4.45	5	46	3336
Bayesian changepoint	0	0	2	6.21	8	45	6081

4 Application to Militarized Interstate Disputes

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 are positively associated with MID initiation. The results are important because they establish the actual empirical association between these variables, whereas theoretically relationships in varying directions are plausible. Moreover, applying the Shewhart chart introduces the question: "Under what conditions are alliances more versus less effective deterrents?," rather than solely focusing on whether alliances appear effective on average.

In order to test for the number of changes and their location, we obtained coefficient estimates over time by fitting a varying-coefficient model across temporal subsets of the dataset. Specifically, we produced estimates by: selecting twenty years of data,[14] fitting the model, storing results, moving forward one year, fitting the model again, and so on.[15] This approach allowed us

14 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.

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

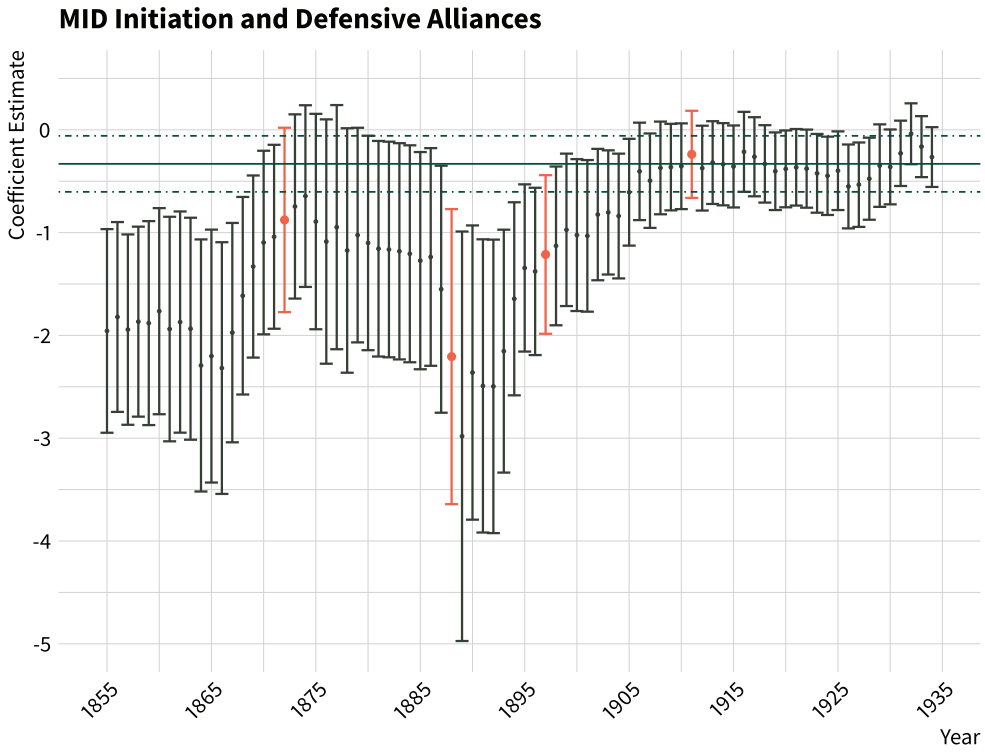

Figure 4. Replication and extension of Leeds (2003). Red, larger estimates correspond to changes detected by the Shewhart chart. Black estimates are considered stable by both techniques. The pooled model's results are a reference point across the figure in green, with dot-dash lines for the pooled confidence interval.

to maximize the number of coefficient estimates while ensuring each model included sufficient data.[16] The results can be seen in Figure 4. First, as a reference point for the previously unmodeled heterogeneity, the pooled model's point-estimate and confidence interval are included across the figure in green with dot-dashed lines for confidence intervals. Next, coefficient estimates which are not detected as a change are included in black and detected changes by the Shewhart chart are in red with larger dots. The estimates which are predicted to be changepoints include data from 1862 to 1882, 1878 to 1898, 1897 to 1907, and 1901 to 1921. It then follows to ask why these time periods correspond with estimated changes, whereas others do not. A full investigation of the substantive sources of these changepoints—and why other time periods are not changes—is beyond the scope of this manuscript. The first two changepoints, however, lend credence to the impact of the Wars of German Unification and then the loss of Bismarck's leadership. The final two changepoints are likely indicative of the move toward World War I and then its start.

Lastly, considering the missed variation that follows from a pooled model, during the long peace associated with the Concert of Europe defensive alliances were associated with far greater deterrent effects than the pooled model finds. This relationship dissipates and almost disappears with WWI, the interwar period, and WWII.[17] This suggests that the pooled estimate is actually somewhat conservative for much of the data.

16 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.

17 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 (e.g., Cranmer and Desmarais 2016).

5 Discussion and Conclusion

We develop a statistical procedure for detecting the locations of coefficient changepoints when fitting generalized linear models to TSCS data. The technique first estimates the bounds for stable statistical behavior through a permutation procedure, which then determines the size of a Shewhart chart that returns the probability of effects changing at each point in time. In a Monte Carlo simulation we find that this procedure outperforms BCP analysis, non-BCP analysis, and CUSUM tests in most simulated conditions. Last, we apply our techniques to a popular study on the relationship between alliances and the onset of conflict, revealing substantial unmodeled effect heterogeneity.

Before concluding, two cautionary points are merited. First, as Table 1 and Figure 2 demonstrate, as the number of changepoints increase in simulation iterations, the more inaccurate all methods become. Given that the exercise is unsupervised, it is not necessarily surprising that all methods struggle with these hard cases. But the results do suggest that one should consider how dynamic the time series of interest appears to be before making strong claims about detected changepoints. If the time series is highly dynamic, then, regardless of method, detecting all changepoints is a substantial statistical challenge. In such a case, one may be better suited investigating the accuracy of their coefficient estimates, exploring potential unmodeled confounders, and more.

Second, while the proposed technique and R package[18] are readily extendable to other statistics of interest that occur over time, one needs the underlying data to apply our technique. More specifically, the null distribution of time-invariant effects can only be estimated if the data (X_{it}, Y_{it}) is available for each randomly selected set of time points. If one solely has access to a time series but none of the data used to generate that time series, then our method cannot be applied. In this case, other methods must be considered.

Although unmodeled effect heterogeneity poses the risk of faulty pooled inferences, it also presents an opportunity to theorize the sources of temporal effect variation. Why might effects change at one time and not another? Why might one change in effects be larger than others? These types of questions are substantively interesting and valuable. However, before these questions can be asked one must differentiate stable variation from genuine changes, which is a statistical challenge. Our proposed technique provides a straightforward means of doing so and improves upon existing approaches.

Data Availability Statement

Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code, and can be viewed interactively at Kent *et al.* (2020a). A preservation copy of the same code and data can also be accessed via Harvard Dataverse at Kent *et al.* (2020b).

References

Aminikhanghahi, S., and D. J. Cook. 2017. "A Survey of Methods for Time Series Change Point Detection." *Knowledge and Information Systems* 51(2):339–367.

Barry, D., and J. A. Hartigan. 1993. "A Bayesian Analysis for Change Point Problems." *Journal of the American Statistical Association* 88(421):309–319.

Beck, N. 1983. "Time-Varying Parameter Regression Models." *American Journal of Political Science* 27:557–600.

Beck, N. 2001. "Time-Series–Cross-Section Data: What Have We Learned In The Past Few Years?" *Annual Review of Political Science* 4(1):271–293.

Beck, N. 2008. "Time-Series-Cross-Section Methods." In *Oxford Handbook of Political Methodology*, edited by J. M. Box-Steffensmeier, H. E. Brady, and D. Collier, 475–493. Oxford: Oxford University Press.

18 https://github.com/dnkent/dynamr

Beck, N., J. N. Katz, and R. Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42:1260–1288.
Blackwell, M. 2018. "Game Changers: Detecting Shifts in Overdispersed Count Data." *Political Analysis* 26(2):230–239.
Box-Steffensmeier, J. M., J. R. Freeman, M. P. Hitt, and J. C. Pevehouse. 2014. *Time Series Analysis for the Social Sciences*. Cambridge, UK: Cambridge University Press.
Box-Steffensmeier, J. M., and B. S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge, UK: Cambridge University Press.
Braumoeller, B. F. 2013. *The Great Powers and the International System: Systemic Theory in Empirical Perspective*. Cambridge, UK: Cambridge University Press.
Carter, D. B., and C. S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18(3):271–292.
Cranmer, S. J., and B. A. Desmarais. 2016. "A Critique of Dyadic Design." *International Studies Quarterly* 60(2):355–362.
Cranmer, S. J., T. Heinrich, and B. A. Desmarais. 2014. "Reciprocity and the Structural Determinants of the International Sanctions Network." *Social Networks* 36:5–22.
De Boef, S., and L. Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52(1):184–200.
Erdman, C., and J. W. Emerson. 2007. "bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems." *Journal of Statistical Software* 23(3):1–13.
Ge, Z., Z. Song, and F. Gao. 2013. "Review of Recent Research on Data-Based Process Monitoring." *Industrial & Engineering Chemistry Research* 52(10):3543–3562.
Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel Hierarchical Models*, vol. 1. New York: Cambridge University Press.
Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach*, vol. 20. Boca Raton, FL: CRC Press.
Golub, J. 2008. "Survival Analysis." In *The Oxford Handbook of Political Methodology*. Oxford, UK: Oxford University Press.
Hastie, T., and R. Tibshirani. 1993. "Varying-Coefficient Models." *Journal of the Royal Statistical Society. Series B (Methodological)* 55:757–796.
Haynes, K., I. A. Eckley, and P. Fearnhead. 2017. "Computationally Efficient Changepoint Detection for a Range of Penalties." *Journal of Computational and Graphical Statistics* 26(1):134–143.
Huckfeldt, R. R., C. W. Kohfeld, and T. W. Likens. 1982. *Dynamic Modeling: An Introduction*. New York: Sage.
Jenke, L., and C. Gelpi. 2016. "Theme and Variations: Historical Contingencies in the Causal Model of Interstate Conflict." *Journal of Conflict Resolution* 63(7): 2262–2284.
Kent, D., J. D. Wilson, and S. J. Cranmer. 2020a. "A Permutation-Based Changepoint Technique for Monitoring Effect Sizes." Code Ocean. https://doi.org/10.24433/CO.5350515.v1.
Kent, D., J. D. Wilson, and S. J. Cranmer. 2020b. "Replication Data for: A Permutation-Based Changepoint Technique for Monitoring Effect Sizes." https://doi.org/10.7910/DVN/NS91FD, Harvard Dataverse, V1, UNF:6:Yww4QlNWYChGp21XaM1P2g== [fileUNF].
Killick, R., and I. Eckley. 2014. "Changepoint: An R Package for Changepoint Analysis." *Journal of Statistical Software* 58(3):1–19.
King, G. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.
Leeds, B. A. 2003. "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.
Mitchell, S. M., S. Gates, and H. Hegre. 1999. "Evolution in Democracy-War Dynamics." *Journal of Conflict Resolution* 43(6):771–792.
Montgomery, D. C. 2013. *Introduction to Statistical Quality Control*. 7th ed. New York: John Wiley and Sons, Inc..
Montgomery, D. C., and J. B. Keats. 1991. *Statistical Process Control in Manufacturing*. New York: Marcel Dekker.
Nieman, M. D. 2016. "Moments in Time: Temporal Patterns in the Effect of Democracy and Trade on Conflict." *Conflict Management and Peace Science* 33(3):273–293.
Park, J. H. 2012. "A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models." *American Journal of Political Science* 56(4):1040–1054.
Thurner, P. W., C. S. Schmid, S. J. Cranmer, and G. Kauermann. 2018. "Network Interdependencies and the Evolution of the International Arms Trade." *Journal of Conflict Resolution* 63(7): 1736–1764.
Wawro, G. J., and I. Katznelson. 2014. "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.
Wilson, J. D., N. T. Stevens, and W. H. Woodall. 2019. "Modeling and Detecting Change in Temporal Networks via the Degree Corrected Stochastic Block Model." *Quality and Reliability Engineering International* 35(5):1363–1378.

Woodall, W. H., and D. C. Montgomery. 1999. "Research Issues and Ideas in Statistical Process Control." *Journal of Quality Technology* 31(4):376–386.
Woodall, W. H., M. J. Zhao, K. Paynabar, R. Sparks, and J. D. Wilson. 2017. "An Overview and Perspective on Social Network Monitoring." *IISE Transactions* 49(3):354–365.
Zorn, C. J. 2001. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." *American Journal of Political Science* 45:470–490.