Theory Testing in the Presence of Causal Heterogeneity*

Nuannuan Xiang†

March 23, 2020

Abstract

Many theories of social science are not universally applicable, but social scientists often test whether a theory holds universally across many cases and over a long time period. Existing methods dealing with this problem either model heterogeneity across units or over time. This paper proposes a new method to model heterogeneity along the unit and time dimensions *simultaneously* with time-series cross-sectional (TSCS) data. Through imposing a sticky hierarchical Dirichlet process hidden Markov model on top of a Dirichlet process mixture model, the proposed method separates time periods into different time regimes and within each regime, categorizes units into different groups. The numbers of time regimes and groups are automatically estimated. The paper illustrates the method by reanalyzing the TSCS data Dunning (2004) uses in studying the relationship between Western aid and democracy in Africa. It automatically detects the end of the Cold War as the watershed, confirming the argument in the original paper, yet identifies a more nature dividing point (1989-1990) than the original paper (1986-1987). It also finds the positive relationship between foreign aid and democracy in the post-Cold War period is driven by a small group of countries, many of which were allies of the Soviet Union during the Cold War period.

*An early version of this paper was written for my methodology preliminary exam. I thank my exam committee members, Kevin Quinn, Yuki Shiraito, Nahomi Ichino, and Nicholas Valentino, for their detailed feedback and discussions. Yuki Shiraito suggested using the sticky HDP-HMM model, which helps improve the paper a lot. I am particularly grateful to Kevin Quinn for providing support and detailed comments and suggestions at every stage of developing this paper. I also thank Fiona Shen-Bayh, Yuequan Guo, Justin Heck, ByungKoo Kim, Saki Kuzushima, Guoer Liu, and Jerry Yu for their helpful discussions. Of course, I alone am responsible for all remaining mistakes and problems.

†Ph.D. candidate, Department of Political Science, University of Michigan. Email: `nnxiang@umich.edu`.

1 Introduction

Theories of social science are often not universally applicable. Many social scientists are interested in studying aggregate entities (countries, states, counties, cities, or other administrative units) over a long time period. For this kind of research, causal relationships often change from one place to another and from time to time. For instance, the classical theory of war and state making derived from European history (Tilly 1990; Downing 1992; Mann 1988) cannot explain state development in Africa (Herbst 1990), or in Latin America (Centeno 2002). The consequential role that civil society played in “making democracy work” in postwar southern Italy (Putnam 1993) and Tocqueville’s America was not replicated in interwar Germany (Berman 1997; Satyanath, Voigtländer and Voth 2017). The resource curse theory seems to hold in the Middle East, but only since the 1970s; other oil-rich countries like Norway, Canada, and Great Britain had few ill effects (Ross 2012). While many studies find that ethnic divisions undermine public good provision (Alesina, Baqir and Easterly 1999; Baldwin and Huber 2010; Miguell and Gugerty 2005), Charnysh (2019) shows that the negative relationship does not exist in migration communities in postwar Poland. And finally, as I will show in the motivating example, depending on what sample researchers use, the effect of foreign on democracy could be positive, negative, or zero.

However, most social scientists test whether a theory holds universally across units and over time. In quantitative analyses, researches often estimate the average causal effects to see whether its distribution follows the implication of the theory they test (usually whether a confidence interval covers 0 or not). Important as the average is, it reveals limited information on whether a theory holds in a specific unit or in a specific time period. Such information is often important for studies in which the unit of analysis is an aggregate entity over a long time period.

Aware of potential causal heterogeneity, some researchers examine how a causal relationship changes with other factors. For instance, Singh and vom Hau (2016) show that the relationship between ethnic diversity and public goods provision changes with different historical trajectories of nation-state formation. For the motivating example of the relationship between foreign aid and democracy, Dunning (2004) argues that the end of the Cold War marked a watershed in the politics of foreign aid in Africa. Correspondingly, in quantitative analyses, researchers are interested in examining how causal effects change with moderator variables. Some researchers use the information of moderator variables to separate their sample into different groups and estimate average effects for each group separately; some researchers employ linear regression models with multiplicative interaction between a moderator variable and the key independent variable (for an improved method, see Hainmueller, Mummolo and Xu 2019); there are also nonparametric approaches to the estimation of heterogeneous effects (Crump et al. 2008; Wager and Athey 2018; Athey and Imbens 2016; Green and Kern 2012). However, all these method require that researches know what moderator variables affect the causal relationship and the data is available.[1]

When information on moderator variables is not available, researchers try to infer the information from data. Some researchers estimate latent group membership to capture heterogeneity across units. Commonly used methods are K-means clustering (Bonhomme and Manresa 2015), Bayesian finite mixture models (Imai and Tingley 2012; Shahn and Madigan 2017), and Dirichlet process mixture models (Shiraito 2016; Ferrari 2020). Some researchers try to detect transition points over time, assuming causal relationships change when entering into a new time regime. Hidden Markov Models (HMM) are often used to estimate transition points (Pang et al. 2012; Park 2011; Brandt and Sandler 2010). Useful as these approaches are, they only examine heterogeneity in one dimension(either in the unit dimension or in the time dimension). For a research project studying many cases over a long time, a method that approaches heterogeneity from both dimensions is in need.

This paper develops a hierarchical Bayesian model to examine causal heterogeneity across units and over time *simultaneously* with time-series cross-sectional (TSCS) data. In short, it contains three layers of analysis. First, in the top layer, it estimates latent time regimes that capture heterogeneity over time. In the middle layer, within each time regime, it estimates latent group membership that captures heterogeneity across units. Third, in the bottom layer, it estimates the average effect of units by group. The model estimates the numbers of time regimes and groups automatically.

For the bottom layer, I use a linear model to estimate the group average effect, but the method can be easily generalized for nonlinear models. For the middle layer, I use a Dirichlet

[1] Or at least researchers have a list of potential moderator variables. Then researchers can employ variable selection methods to detect effective moderators (for example, Wager and Athey 2018; Imai and Strauss 2011).

process mixture model (for example, Hannah, Blei and Powell 2011; Escobar and West 1995) to estimate latent group membership of units. Unlike other clustering approaches such as K-means and Bayesian finite mixture models, Dirichlet process mixture models do not require the number of groups to be prespecified. For the top layer, I use the sticky hierarchical Dirichlet process hidden Markov Model (HDP-HMM) proposed by Fox et al. (2011) to estimate latent time regimes. The sticky HDP-HMM is a revision of the HDP-HMM of Teh et al. (2006), while HDP-HMM is a revision of the classical HMM (for examples of classical HMM, see Barry and Hartigan 1993; Chib 1998). The classical HMM requires the number of time regimes to be prespecified. The HDP-HMM replaces the finite mixture models underlying the classical HMM with a hierarchical Dirichlet process mixture so that the number of time regimes can be estimated automatically. One disadvantage of the HDP-HMM is that it inadequately models the temporal persistence. The sticky HDP-HMM deals with this problem by introducing a "sticky" parameter for self-transition. The sticky HDP-HMM fits social science applicants well, since two adjacent years, for example, are more likely to be in the same time regime than two nonadjacent years.

In the application example, I reanalyze the TSCS data Dunning (2004) uses in the study of Western aid in sub-Saharan African countries. The new result confirms Dunning's argument that the Cold War is the watershed. While Dunning uses domain knowledge to justify 1986-1987 as the dividing point, the new model automatically detects 1989-1990 as the dividing point, which seems to be more natural. Additionally, Dunning argues that after the Cold War donor countries changed to prioritize the goal of promoting democratization over geopolitical calculation, but I find this theory does not hold uniformly across aid-receiving countries. The positive effect of foreign aid is driven by a small group of aid-receiving countries. Many of them are allies of the Soviet Union during the Cold War.

In the following sections, I first introduce this motivating empirical example and the TSCS data in Dunning (2004). I then present the setup of the new model. After this, I conduct simulation studies and apply the model to the empirical example. In the conclusion part, I discuss how to avoid data mining and how to integrate the new model into the research circle of theory testing and theory generating.

2 A Motivating Example: Foreign Aid and Democracy

Does Western aid promote or hinder democracy? The relationship between Western aid and democracy is still under debate. Some studies find that the effect of Western aid on democracy is positive, supporting the hypothesis that Western economic aid is conditioned on liberal reforms (Dietrich and Wright 2015; Bermeo 2016; Carnegie and Marinov 2017). Some studies find that the effect is negative, supporting the hypothesis that aid provides authoritarian incumbents with more resources to cultivate supporters and repress the opposition(Jablonski 2014; Djankov, Montalvo and Reynal-Querol 2008). Other studies find no effect of Western aid on democracy(Hook 1998; Knack 2004).

Dunning (2004) argues that the effect of foreign aid on democracy is different in different international systems. During the Cold War, the Soviet Union and Western countries vied for influence in Africa. With the threat from the Soviet Union, Western donors had to prioritize the goal of cultivating African allies over the goal of promoting democracy, so they were tolerant of aid recipients' nondemocratic institutions. Hence, during the Cold War, Western aid did not promote democracy. After the Cold War when the threat from the Soviet Union disappeared, the geopolitical importance of African allies diminished. Western donors changed to prioritize the goal of promoting democracy. They were more likely to revoke economic aid if aid-receiving countries failed to introduce democratic reforms. Hence, after the Cold War, Western aid promoted democracy.

In the empirical analysis, Dunning uses a TSCS dataset covering 48 sub-Saharan African countries between 1975 and 1997. Democracy is measured by Freedom House "Index of Political Freedom" scores[2]; Western aid is measure by the ratio of official development assistance (ODA) to gross national product (GNP).[3] To test the hypothesis that the end of the Cold War is the watershed, Dunning runs separate linear regressions for the Cold War period (1975-1986) and the post-Cold War period (1987-1997).[4] He finds from 1975 to 1986, there is no statistically significant

[2]The Index has been re-coded so that higher numbers indicate greater political freedom.

[3]Other control variables are a dummy variable for Soviet ally, a dummy variable for common law system, ethnic fractionalization, and GDP per capita.

[4]In the paper, Dunning uses a two-stage least squares (2SLS) model to correct for possible endogeneity. He also runs ordinary least squares (OLS) estimators in supplementary reported, and the results were basically the same as the results of 2SLS.

relationship between ODA and democracy and from 1987 to 1997, the relationship is significantly positive.

1986-1987 seems like an unnatural dividing point. Dunning explains that in the context of Cold War battles for African allies, "heavy Soviet engagement in Africa had already waned by the mid-1980s". However, Dunning also acknowledges that countries like the US started to prioritize pro-democracy agenda only after the early 1990s. Although Dunning uses a Chow test to prove that there is indeed a structural break after 1986, the same test would also support hypotheses of other transition points such as any other years around 1986. Unlike the Chow test, the new model proposed by this paper does not require specifying transition points beforehand. The model estimates transition points to test whether the results are consistent with the hypothesis that the end of the Cold War is the watershed.

Moreover, while Dunning considers causal heterogeneity over time, he doesn't consider causal heterogeneity across aid-receipting countries. For instance, (former) Soviet allies and Western allies might react differently to economic aid from Western donors.[5] Through examining heterogeneity across countries as well as over time, the new model tests not only whether the effect of aid on democracy is universally positive over time, but also whether the effect is universally positive cross countries. The model also facilitates exploring new factors that moderate the effect of Western aid on democracy.[6]

3 The Proposed Methodology

As I have briefly introduced in the beginning, the new method cuts TSCS data into different time regimes; within each time regime, it categorizes units into different groups; within each group, it estimates the group average effect. Correspondingly, it contains three layers of analysis. The top layer is a sticky HDP-HMM to model heterogeneity over time; the middle layer is a Dirichlet process model to model heterogeneity across units; the bottom layer is a linear model to model the causal relationship to be tested. I will explain the three layers from bottom to top, which is also from the simplest to the most complicated.

[5] While Dunning includes a dummy variable for Soviet allies, this does not model the conditional effect of Western aid moderated by the dummy variable of being a Soviet ally.

[6] For existing studies of factors moderating the relationship between foreign and democracy, see Wright 2009; Kono and Montinola 2009; Bermeo 2011; Kersting and Kilby 2014.

3.1 The Bottom Layer: A Bayesian Linear Model

Consider a TSCS dataset of N units over T time periods. Let Y_{it} be the outcome of interest for unit i at time t for $i \in \{1, ..., N\}$ and $t \in \{0, 1, ..., T\}$. Let X_{it} be a vector of explanatory variables including the key independent variable. In this paper, I illustrate the new method by assuming a linear relationship between X_{it} and Y_{it}. The method can be easily generalized for nonlinear models such as generalized linear models.

$$\begin{aligned} Y_{it} &= X_{it}^{\top}\theta_{R(t)G_{R(t)}(i)} + \varepsilon_{it} \\ \varepsilon_{it} &\overset{indep.}{\sim} \mathcal{N}(0, \sigma^2_{R(t)G_{R(t)}(i)}) \end{aligned} \tag{1}$$

$R(t)$ denotes a function that assigns time t to a time regime. $G_{R(t)}(i)$ denotes a function that assigns unit i to a group. The regime assignment $R(t)$ indexes the group assignment $G(i)$, meaning that group membership can be different in different time regimes.

For each time regime $R(t) = r$ with $r \in \{1, 2, ..., \infty\}$ and each group $G_r(i) = g$ with $g \in \{1, 2, 3, ..., \infty\}$ within the time regime r, the prior distributions for θ_{rg} and σ^2_{rg} are:

$$\theta_{rg} \overset{i.i.d}{\sim} \mathcal{N}(v, V) \tag{2}$$

$$\sigma^2_{rg} \overset{i.i.d}{\sim} \text{Inv-Gamma}(\frac{c}{2}, \frac{d}{2}) \tag{3}$$

where v, V, c, d are prior parameters that should be specified beforehand.

If the time regime assignment r and the group assignment g are known, θ_{rg} and σ^2_{rg} can be easily estimated as parameters of a Bayesian linear model. In the following parts, I will introduce how to set up the prior for the regime assignment $R(t)$ and the prior for the group assignment $G_{R(t)}(i)$ so that they can also be conveniently estimated.

3.2 The Middle Layer: A Dirichlet Process Mixture Model

For each time regime $R(t) = r$, I use a Dirichlet process prior for the distribution of group assignment $G_r(i)$ within the time regime r. Specifically, the prior is generated through a stick-breaking process.

$$G_r(i) \overset{i.i.d}{\sim} \text{Discrete}(\{p_g\}_{g=1}^{\infty}) \tag{4}$$

$$p_g = p'_g \prod_{l=1}^{g-1}(1 - p_l) \tag{5}$$

$$p'_g \sim \text{Beta}(1, \alpha) \tag{6}$$

where α is the prior parameter.

Equation (4) indicates that $G_r(i)$ can be regarded as following a discrete distribution with infinite groups. Equation (5) and equation (6) show how to generate the probability for each group. The process of generating the probabilities is like breaking a stick.

Imagine there is a stick of length 1. It will be broken into infinite pieces so that the length of each piece will be the probability of each group. The length of the first piece p_1 simply follows a beta distribution with 1 and the prior parameter α being the two parameters as shown in equation (6). After chopping off the first piece, break the left part to get the second piece. Again, the proportion to be chopped off p'_2 follows a beta distribution shown in equation (6). Then, the length of the second piece p_2 equals the multiplication of the length of the left part $1 - p_1$ and the proportion chopped off p'_2, that is, $p_2 = p'_2(1 - p_1)$ as shown in equation 5. Repeat the same procedure to p_3, p_4, ..., and so on.

The stick-breaking process allows for generating an infinite number of groups for the prior distribution. The probability of a new group p_g converges to zero as the number of groups increases. As a result, in the posterior distribution, the probabilities of most groups become zero and thus, the number of groups can be estimated automatically. The prior parameter α controls the speed of p_g converging to zero. A smaller value of α, which means a larger proportion to be chopped off each time, leaves less space for later pieces, and hence results in faster convergence to zero in the prior distribution. Accordingly, this may lead to a smaller number of groups in the posterior distribution. Conversely, a larger value of α leaves more space for later pieces and hence may result in a larger number of groups in the posterior distribution. The model will be less sensitive to the setting of α as the number of observations increases.

3.3 The Top Layer: A Sticky HDP-HMM

In the top layer of the model, I use the sticky HDP-HMM to model heterogeneity over time. As indicated by the name, the sticky HDP-HMM Model is a special form of hidden Markov model (HMM), meaning the latent time regime $R(t)$ evolves as a discrete-time discrete-state Markov process. In other words, the distribution of $R(t)$ is dependent on $R(t-1)$, the regime assignment in the former period.

Let $\mathbf{\Pi}$ be the transition matrix that regulates the evolvement of a Markov process, with π_{jk} representing the probability of transiting from regime j to regime k. Each row of the transition matrix $\mathbf{\Pi}$ represents the conditional distribution of $P(R(t)|R(t-1)=j)$ with $\sum_{k=1}^{\infty}\pi_{jk}=1$.

$$P(R(t)=k|R(t-1)=j)=\pi_{jk} \tag{7}$$

$$\mathbf{\Pi}=\begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \cdots & \cdots \\ \pi_{21} & \pi_{22} & \pi_{23} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \pi_{j1} & \pi_{j2} & \cdots & \pi_{jk} & \cdots \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \tag{8}$$

where $k=1,2,...,\infty$ and $j=1,2,...,\infty$.

Unlikely the classical HMM, which has a finite number of time regimes (or Markov states) predetermined, the number of time regimes is infinite in the setup of $\mathbf{\Pi}$ here. Also, in the classical HMM, the transition matrix only allows for one-step ahead transition, meaning that R(t) can only stay at the current value or jump to the next higher value. The setup of $\mathbf{\Pi}$ here is more flexible to allow for jumping forward and backward to any regimes. This feature fits applicants of social science better. For example, it is possible that the relationship between Western aid and democracy returns to a stage similar to that during the Cold War period after new competitors like China started to vie for influence in Africa.

In the following, I will explain how to generate π_{jk}, values of the transition matrix $\mathbf{\Pi}$, in detail. First, I will introduce the way of setting up $\mathbf{\Pi}$ in the traditional HDP-HMM, which lays the foundation for the sticky HDP-HMM. Then, I will show how the sticky HDP-HMM improves the traditional HDP-HMM by making two consecutive time periods more likely to be in the same time regime.

HDP-HMM. Define $\pi_{\mathbf{j}} \equiv [\pi_{j1}, \pi_{j2}, ...]$ to represent each row of the transition matrix $\mathbf{\Pi}$. For each starting time regime j, the goal is to generate the probabilities of transiting to a variety of targeting regimes k for $k \in \{1, 2, 3, ..., \infty\}$, with $\sum_{k=1}^{\infty} \pi_{jk} = 1$.

How to generate values of $\pi_{\mathbf{j}}$? In equations (5) and (6), a Dirichlet process has been used to generate a sequence of probability values for a discrete distribution with a infinite number of components. A similar procedure may be used to generate $\pi_{\mathbf{j}}$. However, if $\pi_{\mathbf{1}}$, $\pi_{\mathbf{2}}$, ..., $\pi_{\mathbf{j}}$,... are generated from separate Dirichlet processes, they don't share labels of target time regimes. For instance, the target regime labeled 1 for the starting regime of $j = 1$ does not necessarily has the same feature as the target regime that is also labeled 1 for $j = 2, 3, ...$ One solution is to make $\pi_{\mathbf{1}}$, $\pi_{\mathbf{2}}$, ..., $\pi_{\mathbf{j}}$,... have the same values, the values that generated from one Dirichlet process. However, this arrangement is not flexible enough to capture the difference across different starting time regimes. For example, π_{12}, the probability of moving forward from the starting regime 1 to the target regime 2, may be very different from π_{52}, the probability of jumping backward from the staring regime 5 to the target regime 2 .

The HDP-HMM solves the problem through introducing a dependency structure among separate Dirichlet processes. It bridges separate Dirichlet processes through a shared base measure of target time regimes. This allows for different starting regimes j to have the same labeling system of tarting regimes k. At the same time, $\pi_{\mathbf{1}}$, $\pi_{\mathbf{2}}$, ..., are sill generated from separate Dirichlet processes, which allows for different transition probabilities for different starting regimes j. Details are as the following.

First, generate a discrete base measure for target time regimes k, $\{\beta_k\}_{k=1}^{\infty}$, through a stick-breaking process.

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta_l) \tag{9}$$

$$\beta'_k \sim \text{Beta}(1, \gamma) \tag{10}$$

where γ is the prior parameter.

Second, once the base measure $\{\beta_k\}_{k=1}^{\infty}$ is generated, generate $\pi_{\mathbf{j}}$ for each starging regime $j \in 1, 2, ..., \infty$ through another stick-breaking process.

$$\pi_{jk} = \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi_{jl}) \tag{11}$$

$$\pi'_{jk} \sim \text{Beta}(\alpha_0 \beta_k, \alpha_0 \sum_{l=1}^{k} \beta_l) \tag{12}$$

where α_0 is the prior parameter.

Like the classical Dirichlet process, the hierarchical Dirichlet process also generates an infinite number of time regimes (groups) with the probabilities of new time regimes declining quickly as the number of time regimes increases. Hence, in the posterior distribution, probabilities of transition to most time regimes are zero, which enables the model to estimate the number of time regimes automatically. Prior parameters α_0 and γ regulate the concentration of group membership. A larger α_0 or γ may result in more time regimes in the posterior distribution, but when a dataset covers a longer time, the model becomes less sensitive to the setting of α_0 and γ.

3.4 Sticky HDP-HMM

Since the HDP-HMM is developed from the classic Dirichlet process model, which is designed for cross-sectional data, it doesn't consider the situation that two consecutive time periods are more likely to be in the same regime. Ignoring the time structure threatens to over fragment the data and produce redundant time regimes. The sticky HDP-HMM proposed by Fox et al. (2011) deals with this problem by introducing a self-transition bias, which makes two consecutive time periods more likely to be assigned to the same regime in the prior distribution. In detail, replace equation (12) with

$$\pi'_{jk} \sim \text{Beta}((\alpha_0 + \omega)(\frac{\alpha_0 \beta_k + \omega \mathbf{I}(j = k)}{\alpha + \omega}), (\alpha_0 + \omega)(1 - \sum_{l=1}^{k} (\frac{\alpha_0 \beta_l + \omega \mathbf{I}(j = l)}{\alpha + \omega}))) \tag{13}$$

where ω ($\omega \geq 0$) is the prior parameter that introduces self-transition bias .

If $\omega = 0$, equation (13) is exactly the same as equation (12). For positive values of ω, if

$j = k$, which indicates a self-transition from regime j to regime j, the first parameter of the Beta distribution in equation (13) becomes larger and the second parameter becomes smaller, resulting in a larger probability for saying in regime j than transiting to other regimes.The larger the prior parameter ω is, the "stickier" the prior distribution is, which may result in less fragmented data in the posterior distribution.

3.5 Markov Chain Monte Carlo Algorithm for Estimation

The model is estimated by a Gibbs sampling algorithm. Each Markov Chain Monte Carlo iteration includes three major steps. The first step is to update parameters of the linear models θ_{rg} and σ^2_{rg} for each group g in each time regime r. Given current regime and group assignments, the posterior distribution for θ_{rg} is a normal distribution and for σ^2_{rg} is a Inverse Gamma distribution. The second step is to update group assignment $G_r(i)$ for each time regime r. I use a truncated Gibbs sampler (Ishwaran and James 2001) to sample group assignment. Although in theory, the number of groups is infinite in the prior distribution, in practice, the group number can be an arbitrarily large number to approximate the theoretical Dirichlet Process prior. In this way, the posterior distribution of $G_r(i)$ becomes a discrete distribution with limited components, which can be easily sampled. The third step is to update time regime assignment $R(t)$. Again, the number of time regimes is set to an arbitrarily large number to approximate the (hierarchical) Dirichlet prior. I use a forward-backward algorithm to sample R(t) in sequence. Details of the algorithm can be find in the Appendix.

4 Simulation Studies

The dataset I will reanalyze in the application example is an unbalanced TSCS dataset in which some countries have data in the early years missing. Since unbalanced TSCS datasets are common in social science studies, in this section, I design simulation studies to check whether the proposed method works for unbalanced datasets and under what conditions it works (or not works).

I keep the data of the independent variables (`foreign aid, common law, ethnicity fragmentation, GDP per capita`) in the original dataset of Dunning (2004). Using the data of the independent variables, I generate new data of the dependent variable of (`democracy score`) through linear models. In this way, I keep the structure of missing data in the original dataset.

I investigate three major situations: 1) One dividing point that creates two different time regimes, 2) two dividing points that only create two different time regimes, with the transition sequence being regime 1 to regime 2 to regime 1, and 3) two dividing points that create three different time regimes. Within each time regime, I generate 2 or 3 groups. The coefficient for the key independent variable of `foreign aid` varies from group to group.

As expected, the model performs better when the number of groups is small, the number of observations in each group is large, and the coefficients of different groups are not very close to each other. Also, I experiment with a variety of commonly used prior parameters and find the estimation results do not change much with different priors for this dataset.[7] However, like other Bayesian models, estimations with a very small number of units or short time length may be sensitive to prior parameters. Users should always check how the setting of prior parameters influences the results.

4.1 Simulation 1: One Dividing Point, Two Time Regimes

I begin with a simple setting that one dividing point creates two regimes. I set the dividing point to 12-13, which is equivalent to 1986-1987 as in the original paper of Dunning (2005). In the first regime, I generate three groups. The first group includes countries categorized as "Soviet clients" in the original paper. For other countries, I randomly assign them to the second group and the third group. In the second regime, there are only two groups. The first group is also composed of "Soviet clients" and the other group is composed of all other countries left.

Different groups have different coefficients (of linear models) for the independent variables. Details of the coefficients are explained in Table 1. The coefficients are very similar to the results of the original paper, except that the coefficients for the constant and the key independent variable `foreign aid` are set to be different across groups. I use linear models to generate data of the dependent variable based on the information of the coefficients and the data of the independent variables. The standard deviation of the error term is 0.1, the same for all groups.

Then I use the new data of the dependent variable and the original data of the independent

[7] The prior parameters I use are: $v = 0$, $V = \{\text{diag}(5), \text{diag}(10), \text{diag}(20), \text{diag}(50)\}$, $c = \{1, 2\}$, $d = \{1, 2\}$, $\alpha = \{0.5, 1, 1.5, 2\}$, $\alpha_0 = \{0.5, 1, 1.5, 2\}$, $\gamma = \{0.5, 1, 1.5, 2\}$, $\omega = \{0, 0.5, 1, 1.5\}$.

variables [8] to estimate the model. Of course, information on the dividing point, group memberships, as well as the number of time regimes and groups are hidden. They are to be estimated by the model.

The model works well in identifying 12-13 as the only dividing points. As shown in the up-left part of Figure 1, for most of the time points, the probabilities of transiting to a new regime (calculated as the probabilities that the current point and the former point are in two different regimes) are near 0. At time point 12, the probability increases a little bit; at time point 13, the probability is near 1, indicating that point 12 and point 13 are very likely to be in two different regimes. Correspondingly, as shown in the down-left part of Figure 1, the average of the coefficients for the key independent variable (`foreign aid`) changes suddenly after time point 13.

While the model works well in estimating the year-averaged coefficients for regime 2, the model does not recover the true values precisely for regime 1. The reason is to be explored in Figure 2.

Figure 2 shows the distributions of the coefficients for the key independent variable. As the true group memberships are known, I plot the histograms of observations in different groups with different colors. In this way, the plot can reveal the proportions of different groups in a certain distribution range along the x-axis. If the model recovers the true memberships, different colors will be separated; otherwise, the colors will mix together, indicating that the model wrongly assigned similar values to units with different group memberships.

As shown in the up-right part of Figure 2, for regime 2, the model recovers the true membership well. However, for regime 1, there is a small portion of units in group 3 (in which the true coefficient of interest is 0.02) is wrongly assigned to group 2 (in which the true coefficient of interest is 0.01). This is understandable because group 2 and group 3 are not very different from each other (see Table 1). Missing data also makes it more difficult for the model to estimate group memberships, since some units have data in the early years missing, which leaves the model limited information for assigning groups. This also explains why the year-averaged coefficients are not precisely estimated, as shown in the down-left part of Figure 1.

[8] The dummy variable for "Soviet Client" is not included in the independent variables. The group membership of "Soviet Client" is to be estimated by the model.

Table 1: Coefficients of Linear Models in Simulation Studies

Simulation	Group	Regime 1	Regime 2	Regime 3
Simulation 1	Group 1	[-3.5, -0.02, 1, 0, 0.7]	[-3.5, 0.02, 0.4, 0, 1]	NA
	Group 2	[-4.0, 0.00, 1, 0, 0.7]	[-4.0, 0.00, 0.4, 0, 1]	NA
	Group 3	[-4.0, 0.02, 1, 0, 0.7]	NA	NA
Simulation 2	Group 1	[-3.5, -0.02, 1, 0, 0.7]	[-3.5, 0.02, 0.4, 0, 1]	same as Regime 1
	Group 2	[-4.0, 0.00, 1, 0, 0.7]	[-4.0, 0.00, 0.4, 0, 1]	
	Group 3	[-4.0, 0.02, 1, 0, 0.7]	NA	
Simulation 3	Group 1	[-3.5, -0.02, 1, 0, 0.7]	[-3.5, 0.02, 0.4, 0, 1]	[-3.5, 0.04, 0.4,0,1]
	Group 2	[-4.0, 0.00, 1, 0, 0.7]	[-4.0, 0.00, 0.4, 0, 1]	[-4, 0.02 , 0.4,0,1]
	Group 3	[-4.0, 0.02, 1, 0, 0.7]	NA	[-4, -0.02 , 0.4,0,1]

Note: The coefficients are for the constant, `foreign aid`, `common law`, `ethnicity fragmentation`, and `GDP per capita` respectively.

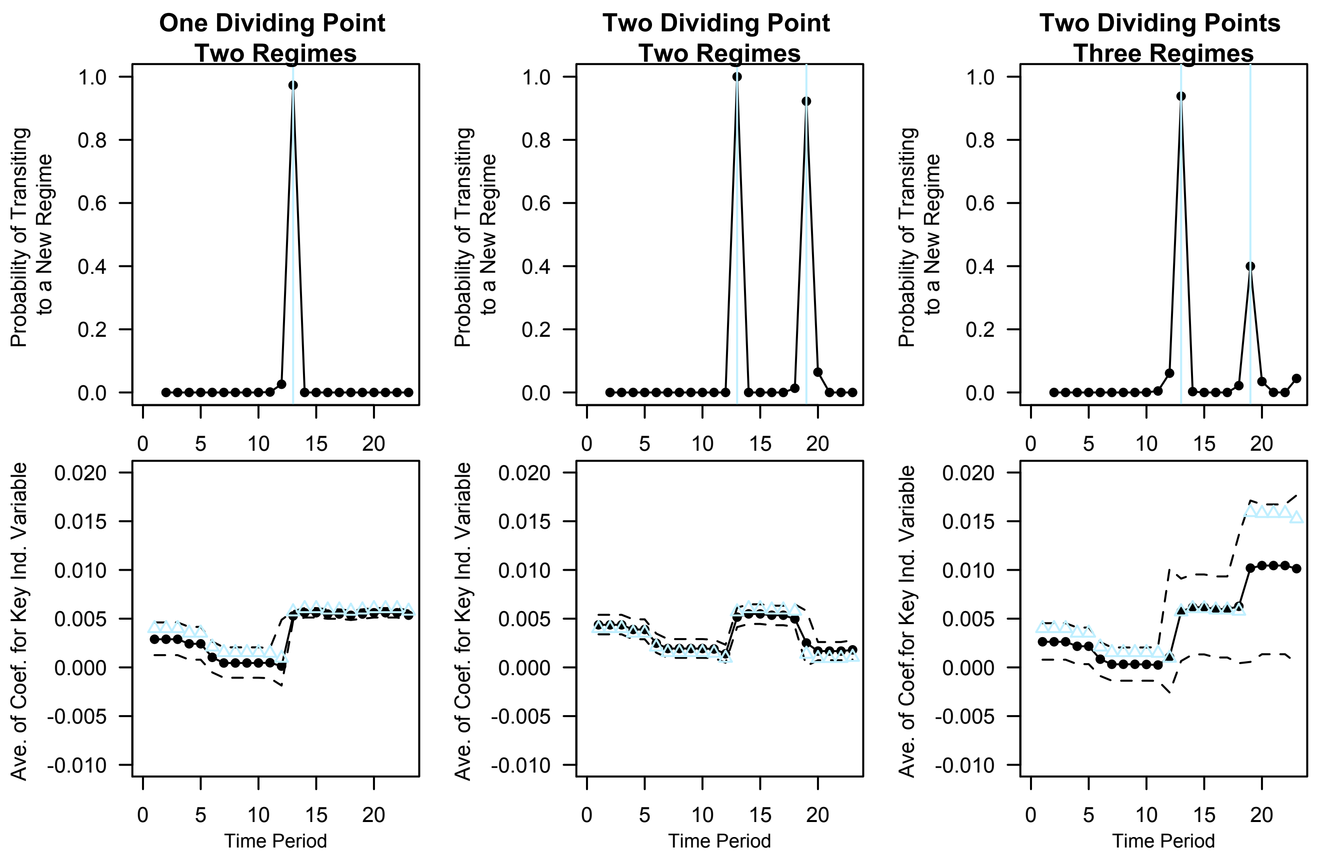
4.2 Simulation 2: Two Dividing Points, Two Time Regimes

In simulation 2, there are two dividing points, one at 12-13 as in simulation 1 and the other at 18-19. The number of time regimes is still two. The first thirteen time periods are in regime 1, the next six periods are in regime 2, and the last five periods return to regime 1. For both time regimes, group memberships are still the same as those in simulation 1, with three groups in regime 1 and two groups in regime 2. For each group, the coefficients that are used to generate data of the dependent variable are also the same as those in simulation 1 (details are in Table 1).

As shown in Figure 1 (the up-middle part), the model works well in identifying the two dividing points. Correspondingly, the year-averaged coefficient for the key independent variable (shown in the down-middle part of Figure 1) increases suddenly at point 13 and goes back to the former level after point 19, indicating a transition back to regime 1.

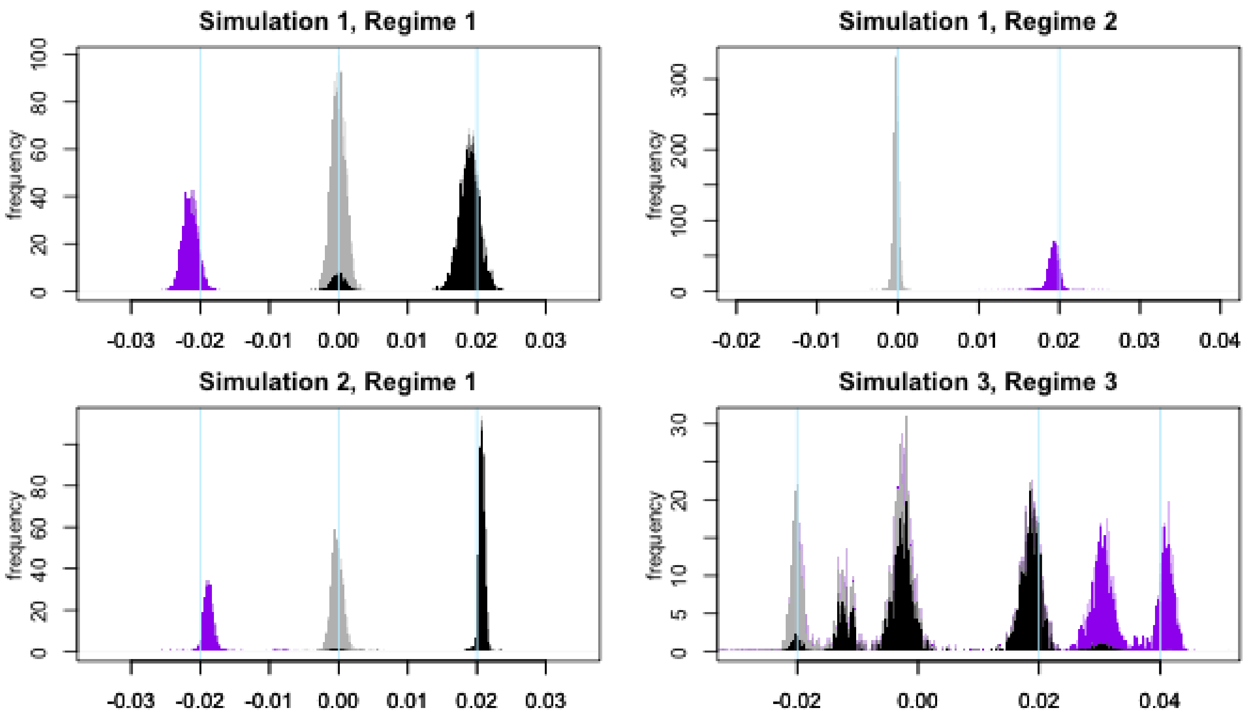
Compared to simulation 1 (the up-left part of Figure 1), the model works even better in identifying the first dividing point. The probability of transiting to a new regime is near 0 at point 12 and near 1 at point 13. Also, the estimation of the year-averaged coefficients for regime 1 is more accurate and the 95% credible intervals become narrower. The distribution of the coefficients (the down-left part of Figure 2) shows that the better-estimated group memberships lead to the better-estimated year-averaged coefficients. Unlike the results of Simulation 1 (the up-left part of Figure 1), few units are assigned to the wrong groups. The distribution is also more concentrated, which results in narrower credible intervals of the year-averaged coefficients.

Figure 1: Estimated Dividing Points and Averages of Coefficients for the Key Independent Variable in the Simulation Studies



Note: The figure reports the results of three simulations. The upper part reports the probabilities that the current point and the former point are in different regimes; the lower part reports the year-average of the coefficients for the key independent variable. While a variety of prior parameters are used to estimate the models, the results do not change much with different prior parameters. In this figure, the prior parameters of $v, V, c, d, \alpha, \alpha_0, \gamma, \omega$ are 0, diag(10), 1, 1, 1, 1, 1, 0.5 . For each estimation, 30,000 Markov Chain Monte Carlo iterations are implemented, with 10,000 discarded burn-in iterations. In the upper part, the blue lines represent the true dividing points; the the lower part, the blue dots represent the true year-average values of the coefficient for the key independent variable. The dashed black lines show a 95% credible interval for the year-average coefficients.

Figure 2: Smoothed Histograms of Estimated Coefficients for the Key Independent Variables in the Simulation Studies, Colored with True Group Memberships



Note: The figure reports the distributions of estimated coefficients for the key independent variables in selected regimes and selected simulations. I use the information of true group memberships to color the plots. A mix-up of colors indicates the model does not recover the true membership. The blue line denotes the true values of the coefficients.In this figure, the prior parameters of $v, V, c, d, \alpha, \alpha_0, \gamma, \omega$ are 0, diag(10), 1, 1, 1, 1, 1, 0.5 . For each estimation, 30,000 Markov Chain Monte Carlo iterations are implemented, with 10,000 discarded burn-in iterations.

Why is the estimation for regime 1 better in simulation 2 than in simulation 1? The reason is that there becomes more information as the last five years are added to regime 1. Particularly, since many units have missing data in the early years, adding the data of the last five years gives the model more information to infer group memberships. Also, with more data, the estimation of the coefficients become more precise. Then, this leads to better inference on the dividing point.

4.3 Simulation 3: Two Dividing Points, Three Time Regimes

Like in simulation 2, there are also two dividing points in simulation 3, with one at 12-13 and the other at 18-19. Unlike simulation 2, there are three regimes in simulation 3. The first two regimes are exactly the same as those in simulation 1 (see Table 1 for details of the coefficients). For the third regime, I also categorize units into three groups. The first group is still composed of units labeled "Soviet clients" in the original paper of Dunning (2004). For other units, I randomly assign them into two groups. [9] To make it more difficult for the model to detect the dividing point between regime 2 and regime 3, I make the coefficients of regime 3 very similar to the coefficients of regime 2. As shown in Table 1, the coefficients of the first group in regime 2 and regime 3 are only slightly different; the average coefficients of the second and the third groups in regime 2 are the same as the coefficients of the second group in regime 3.

The results are shown in the right part of Figure 1. The model can still detect the two dividing points. However, the second dividing point of 18-19 is not as clear as the first one of 12-13. [10] The probability of entering into a new regime at time point 13 is around 0.4, indicating that the model frequently assigned the last 5 time points into regime 2. Correspondingly, the year-averaged coefficients of the key independent variable are not correctly estimated and the 95% credible intervals are very wide for regime 3 (the down-left part of Figure 1). The credible intervals for regime 2 also become wider, as observations in regime 3 are frequently assigned to regime 2 to add noises to the estimation of the coefficients.

I further explore what could go wrong by looking into the distribution of the estimated coefficient in regime 3. As shown in Figure 2 (the down-right part), the number of clusters is more than

[9] It should be noticed that group memberships of regime 1 and regime 3 are different.

[10] I also look into cases in which regime 2 and regime 3 are very different from each other. The model works well in detecting the transition point.

the true number of groups. For units of the first group (with the true coefficient 0.04), they are frequently assigned values around 0.03. This is when regime 3 and regime 2 are wrongly merged and the model samples a value around the average of the true coefficient of 0.04 in regime 3 and 0.02 in regime 2. For the second group (with the true coefficient 0.02) and the third group (with the true coefficient -0.02), they were wrongly mixed and assigned values near 0 when regime 3 is merged into regime 2.

Simulation 3 reveals that the model does not work well when two regimes are very similar to each other. When there are no clear-cut dividing points (the probability of transiting to a new regime is not close to 1), the estimated results of group memberships and the coefficients will be misleading.

5 Application Example: The Effect of Western Aid and Democracy in Africa

In this section, I apply the proposed method to reanalyze the dataset of Dunning (2004). There are three major hypotheses to test. The first hypothesis is that the effect of Western aid on democracy is positive. In this test, I examine whether the general argument that aid promotes democracy holds universally. Second, I examine the argument of Dunning (2004) that the end of the Cold War marked a watershed in the politics of foreign aid in Africa. Third, I exam the other argument of Dunning (2004) that the principle of conditioning Western aid on democratic reforms was only nominal during the Cold War and became substantial after the Cold War. I test the hypothesis that the effect of aid is near zero during the Cold war and becomes positive after the Cold War.

Like the original paper, I assume a linear relationship between the dependent variable `democracy score` and the independent variables of `foreign aid`, `common law`, `ethnicity fragmentation`, and `GDP per capita`.[11] The parameter of interest is the coefficient for the key independent variable of foreign aid. I examine how the effect of foreign aid on democracy changes across units and

[11] I don't include the dummy variable of Soviet client into the independent variables because I will use this application example to illustrate that the model can automatically detect the difference between the Soviet Bloc and the Western Bloc later. The results of the estimation with the variable of Soviet client are not very different from the results presented here. In the original paper, Dunning also finds that the results are robust to the exclusion of the variable of Soviet client. See Dunning (2004), note 23.

over time.

Major findings are presented in Figure 3 and Figure 4.[12] Figure 3 presents the heterogeneity over time. It shows that 1989-1990 is a clear-cut dividing point. In the year of 1990, the probability of transiting to a new time regime is near 1, while the probabilities in all other years are near zero. Consistent with the findings of Dunning (2004), this suggests that Western aid does not promote democracy universally and the end of the Cold War is an important transition point. Also consistent with the results in the original paper, I find that the estimated effect of foreign aid on democracy is near zero during the Cold War period and it becomes larger after the Cold War.[13]

The dividing point of 1989-1990 is different from the dividing point of 1986-1987 of the original paper. Using the domain knowledge, Dunning judges that 1986-1987 should be the dividing point because, by the mid-1980s, Soviet foreign policy dominated by Mikhail Gorbachev, Eduard Shevardnadze, and others had already less stressed the Soviet engagement in Africa. However, for Western countries, the Soviet Union as a constant threat had not disappeared by 1986. It is not very likely that Western countries changed their strategies in Africa immediately after a "new thinking" emerged in Soviet foreign policy. They needed more time to observe and judge the new situation. In this sense, 1989-1990 seems to be a more natural dividing point. Although the Soviet Union did not collapse until 1991, its dissolution had loomed large in 1990, which might motivate Western donors to change their African policies. For instance, in Franco-Africa Summit of 1990, French President Mitterrand announced that future French aid would be linked to democratization (Chafer 2002).[14]

Figure 2 illustrates the heterogeneity across countries. For each time regime, I plot the density of the estimated effect of aid. For the Cold War period, the density concentrates on the value of zero, confirming Dunning's argument that during the Cold War, Western aid was not conditional on implementing democratic reforms. For the post-Cold War period, the major peak of the density moves slightly toward positive values. At the same time, there emerges a small peak around the

[12] I estimate the model with a variety of prior parameters and the results do not change much. The prior parameters I use are: $v = 0$, $V = \{\text{diag}(5), \text{diag}(10), \text{diag}(20), \text{diag}(50)\}$, $c = \{1, 2\}$, $d = \{1, 2\}$, $\alpha = \{0.5, 1, 1.5, 2\}$, $\alpha_0 = \{0.5, 1, 1.5, 2\}$, $\gamma = \{0.5, 1, 1.5, 2\}$, $\omega = \{0, 0.5, 1, 1.5\}$.

[13] Although a 95% credible interval cover zero and negative values, it is reasonable because the variance of the year-averaged effect should be larger when the heterogeneity across units is considered in the model.

[14] I thank Nahomi Ichino for pointing this out.

value of 0.06, which is much larger than the value of 0.012, the result of the original paper. This suggests that the result in the original paper may be driven by a small group of countries.

To investigate how the effect of aid varies from country to country in the Post-Cold War period, I construct 95% credible intervals for each country. Table 2 lists all countries that have credible intervals cover either all positive values or all negative values. It shows that 11 out of 48 countries have 95% credible intervals covering all positive values and 7 out of 48 countries have credible intervals covering all negative values.

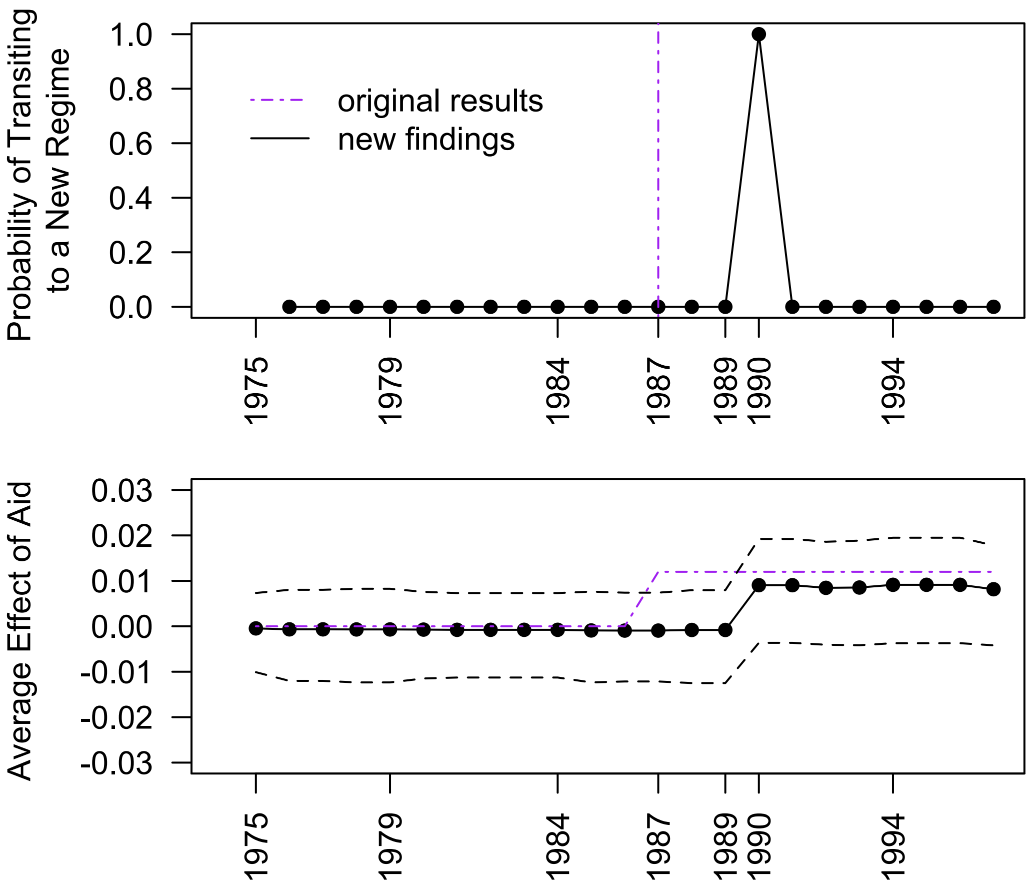
Table 2 also explores what kinds of countries have all positive or negative values. It separates countries into the Soviet bloc and the Western bloc. I code a country as a country in the Soviet bloc if the Soviet Union considered a sub-Saharan African state a "revolutionary democracy" or "socialist-oriented", the same as the way Dunning (2004) codes the variable of Soviet client. I assign all other countries to the Western bloc.[15] There are thirteen countries in the Soviet bloc and 35 countries in the Western bloc.

Table 2 shows that more than half of the countries with all positive values were in the Soviet bloc during the Cold War period. This casts doubt on the positive relationship between foreign aid and democracy in these countries and in general after the Cold War. With the collapse of the Soviet Union, many Soviet allies implemented democratic reforms. At the same time, they started to receive more Western aid. Did Western aid result in democratic reforms, or the two events just happened together? If Western aid indeed promoted democracy in these countries, did Western donors implement the same foreign policies in other aid-receiving countries? Or donors just selected several former Soviet allies to demonstrate the values of free-market economy and democracy? If so, why these countries were selected? To answer these questions, careful studies of causal mechanisms in individual cases are required.

In other words, the results provide a base to explore under what conditions foreign aid promotes or hinders democracy. For instance, more researches may be done to explain why cases like Burundi, Rwanda, and Djibouti have a negative effect of aid on democracy? What common features do these countries share? Exploring these questions may lead to a new argument about the heterogeneous effect of Western aid.

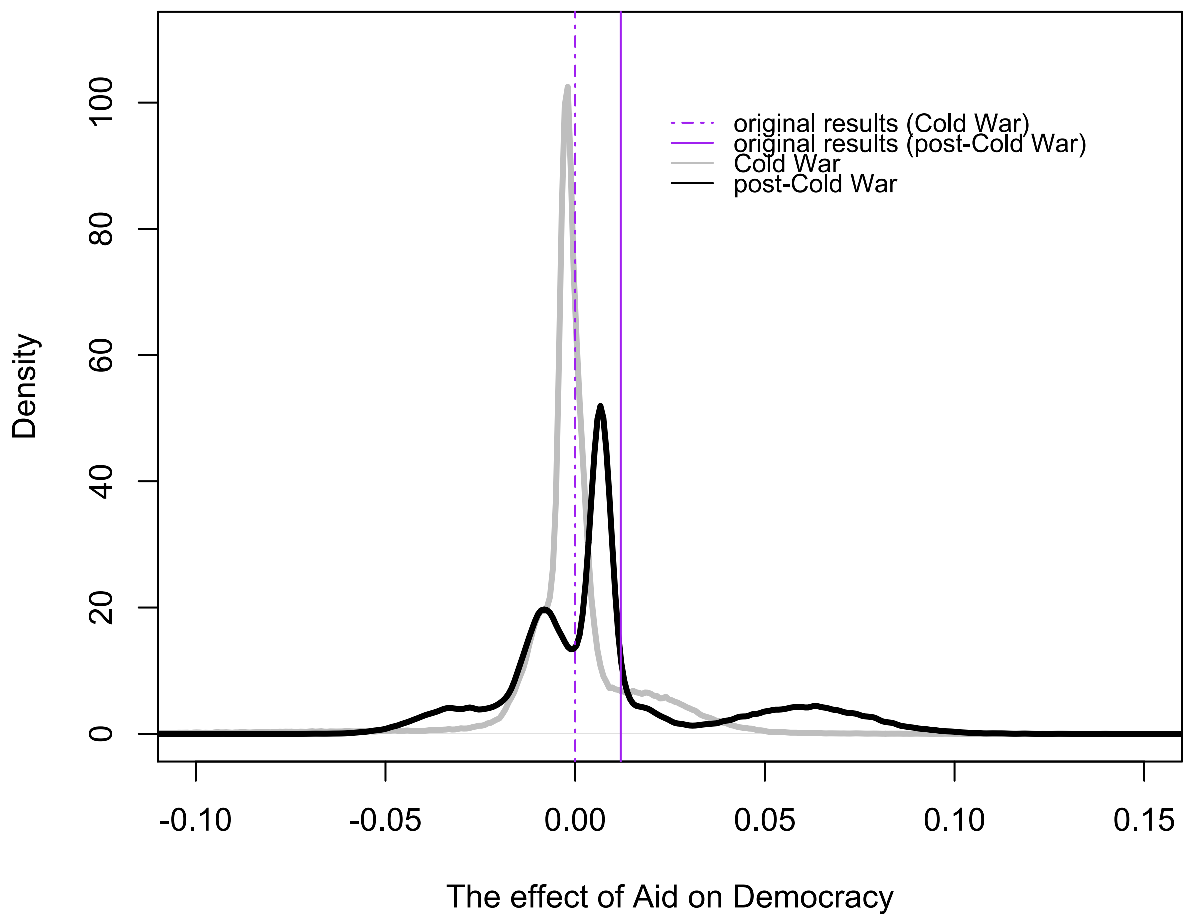
[15]Dunning (2004) adopts the measure from Albright (1991).

Figure 3: Estimated Dividing Points and the Year-averaged Effects of Foreign on Democracy



Note: The prior parameters of $v, V, c, d, \alpha, \alpha_0, \gamma, \omega$ are 0, diag(10), 1, 1, 1, 1, 1, 0.5 . For each estimation, 30,000 Markov Chain Monte Carlo iterations are implemented, with 10,000 discarded burn-in iterations. The purple lines show the results of the original paper of Dunning (2004). The black solid lines and dots are the results estimated with the new method. The dashed black lines show a 95% credible interval for the year-averaged effects of Western aid on democracy.

Figure 4: Smoothed Histograms of the Effects of Western Aid on Democracy in Africa, before and after the Cold War



Note: The prior parameters of $v, V, c, d, \alpha, \alpha_0, \gamma, \omega$ are 0, diag(10), 1, 1, 1, 1, 1, 0.5 . For each estimation, 30,000 Markov Chain Monte Carlo iterations are implemented, with 10,000 discarded burn-in iterations. The purple lines show the results of the original paper of Dunning (2004). The grey line represents the density of the Cold War period and the black line represents the density of the post-Cold War period.

Table 2: Countries with 95% Credible Intervals of the Effect of Aid on Democracy Covering All Positive or Negative Values in the Post-Cold War Period

Blocs	Positive Effects of Aid	Negative Effect of Aid
The Soviet Bloc	(6/13) Benin Cape Verde Madagascar Mali Seychelles Zambia	(1/13) Angola
The Western Bloc	(5/35) The Gambia Ghana Malawi São Tomé and Príncipe South Africa	(6/35) Botswana Burundi Cameroon Djibouti Mauritania Rwanda

Note: Countries that are assigned to the Soviet Bloc are countries countries coded as Soviet clients in Dunning (2004). All other countries are assigned to the Western Bloc.

6 Concluding Remarks

A large-N dataset is often used to test whether a theory holds widely, yet this ignores the importance of using a large-N dataset to examine and explore causal heterogeneity. This paper develops a new model to analyze heterogeneous relationships across units and over time simultaneously. Besides testing whether a theory is universally applicable, it also identifies among which units and in what time periods the theory holds. Thus, it helps explore the scope condition of the theory and facilitates generating new theories.

It should be noticed that the model is not suggesting a new way of data mining: Use the proposed method to search for subgroups with positive or negative effects, extrapolate a new argument about the heterogeneous effect based on the results, and then reuse the results to support the argument. To test an argument about the heterogeneous effect, the researcher must specify in advance the argument. For instance, in the application example, I prespecify the argument that the Cold War is the dividing point. Then I examine whether the results are consistent with this argument. Although I explore what factors differentiate the effect of foreign aid after the Cold

War and finding the factor of being a Soviet ally may be a potential moderator, I cannot use the results of the proposed model as the evidence to support the argument. As I have mentioned above, new evidence is needed to test the argument.

In this sense, the proposed method suggests a new research circle of theory generating and testing. In a traditional research circle, outliers are identified in a large-N analysis, studying important outlier cases helps identify the missing explanatory variables, and then a new large-N analysis is implemented to test the new theory (Lieberman 2005). In the new circle, a large-N analysis with the proposed method helps discover subgroups with different relationships of interest. Studying representative cases of each group helps generate new arguments about moderator factors. Then new data is collected to test the argument in a new large-N analysis. To summarize, the traditional research circle generates and tests theories of new explanatory variables, the new research circle generates and tests theories of new moderator variables.

References

Albright, David E. 1991. "Soviet Economic Development and the Third World." *Soviet Studies* 43(1):27–59.

Alesina, Alberto, Reza Baqir and William Easterly. 1999. "Public Goods and Ethnic Divisions." *The Quarterly Journal of Economics* 114(4):1243–1284.

Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Baldwin, Kate and John D. Huber. 2010. "Economic versus Cultural Differences: Forms of Ethnic Diversity and Public Goods Provision." *American Political Science Review* 104(4):644–662.

Barry, Daniel and J. A. Hartigan. 1993. "A Bayesian Analysis for Change Point Problem." *Journal of the American Statistical Association* 88(309-319).

Berman, Sheri. 1997. "Civil Society and the Collapse of the Weimar Republic." *World Politics* 49(3):401–429.

Bermeo, Sarah Blodgett. 2011. "Foreign Aid and Regime Change: A role for Donor Intent." *World Development* 39(11):2021–2031.

Bermeo, Sarah Blodgett. 2016. "Aid Is Not Oil: Donor Utility, Heterogeneous Aid, and the Aid-Democratization Relationship." *International Organization* 70(1).

Bonhomme, Stéphane and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83(3):1147–1184.

Brandt, Patrick T. and Todd Sandler. 2010. "What Do Transnatioinal Terrorists Target? Has It Changed? Are We Safe?" *Journal of Conflict Resolution* 54:214–236.

Carnegie, Allison and Nikolay Marinov. 2017. "Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment." *American Journal of Political Science* 61(3):671–683.

Centeno, Miguel A. 2002. *Blood and Debt: War and the Nation-state in Latin America*. University Park, Pa.: Pennsylvania State University Press.

Chafer, Tony. 2002. "Franco-African Relations: No Longer so Exceptional?" *African Affairs* 101(404):417–436.

Charnysh, Volha. 2019. "Diversity, Institutions, and Economic Outcomes: Post-WWII Displacement in Poland." *American Political Science Review* .

Chib, Siddhartha. 1998. "Estimation and Comparison of Multiple Change-point Models." *Journal of Econometrics* 86(2):221–241.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens and Oscar A. Mitnik. 2008. "Nonparametric Tests for Treatment Effect Heterogeneity." *The Review of Economics and Statistics* 90(3):389–405.

Dietrich, Simone and Joseph Wright. 2015. "Foreign Aid Allocation Tactics and Democratic Change in Africa." *The Journal of Politics* 77(1):216–234.

Djankov, Simeon, Jose G. Montalvo and Marta Reynal-Querol. 2008. "The Curse of Aid." *Journal of Economic Growth* 13(169-194).

Downing, Brian M. 1992. *The Military Revolution and Political Change: Origins of Democracy and Autocracy in Early Modern Europe*. Princeton: Princeton University Press.

Dunning, Thad. 2004. "Conditioning the Effects of Aid: Cold War Politics, Donor Credibility, and Democracy in Africa." *International Organization* 58(1):409–423.

Escobar, Michael D. and Mike West. 1995. "Bayesian Density Estimation and Inference Using Mixtures." *Maximum Entropy and Bayesian Methods* 90(430):577–588.

Ferrari, Diogo. 2020. "Modeling Context-Dependent Latent Effect Heterogeneity." *Political Analysis* 28(1):20–46. Working paper.

Fox, Emily B., Erik B. Sudderth, Michael I. Jordan and Alan S. Willsky. 2011. "A Sticky HDP-HMM with Application to Speaker Diarization." *The Annals of Applied Statistics* 5(2A):1020–1056.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* .

Hannah, Lauren A., David M. Blei and Warren B. Powell. 2011. "Dirichlet Process Mixtures of Generalized Linear Models." *Journal of Machine Learning Research* 12:1923–1953.

Herbst, Jeffrey. 1990. *States and Power in Africa: Comparative Lessons in Authority and Control*. Princeton and Oxford: Princeton University Press.

Hook, Steven W. 1998. "'Building Democracy' through Foreign Aid: The Limitations of United States Political Conditionalities, 1992–96." *Democratization* 3(156-180).

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.

Ishwaran, Hemant and Lancelot F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96(453):161–172.

Jablonski, Ryan S. 2014. "How Aid Targets Votes: The Impact of Electoral Incentives on Foreign Aid Distribution." *World Politics* 66(2):293–330.

Kersting, Erasmus and Christopher Kilby. 2014. "Aid and Democracy Redux." *European Economic Review* 67(C):125–143.

Knack, Stephen. 2004. "Does Foreign Aid Promote Democracy." *International Studies Quarterly* 48(1):251–266.

Kono, Daniel Yuichi and Gabriella R. Montinola. 2009. "Does Foreign Aid Support Autocrats, Democrats, or Both?" *Journal of Politics* 71(2):704–718.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *The American Political Science Review* 99(3):435–452.

Mann, Michael. 1988. *States, War and Capitalism: Studies in Political Sociology.* Cambridge: Basil Blackwell.

Miguell, Edward and Mary Kay Gugerty. 2005. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Journal of Public Economics* .

Pang, Xun, Barry Friedman, Andrew D. Martin and Kevin M. Quinn. 2012. "Endogenous Jurisprudential Regimes." *Political Analysis* 20(4):417–436.

Park, Jong Hee. 2011. "Change-point Analysis of Binary and Ordinal Probit Models: An Application to Bank Rate Policy under the Interwar Gold Standard." *Political Analysis* 19(2):417–436.

Putnam, Robert D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy.* New Jersey: Princeton University Press.

Ross, Michael L. 2012. *The Oil Curse: How Petroleum Wealth Shapes the Development of Nations.* Princeton: Princeton University Press.

Satyanath, Shanker, Nico Voigtländer and Hans-Joachim Voth. 2017. "Bowling for fascism: Social Capital and the Rise of the Nazi Party." *Journal of Political Economy* 125(2):478–526.

Shahn, Zach and David Madigan. 2017. "Latent Class Mixture Models of Treatment E ect Latent Class Mixture Models of Treatment Effect Heterogeneity." *Bayesian Analysis* 12(3):831–854.

Shiraito, Yuki. 2016. "Uncovering Heterogeneous Treatment Effects." Working paper.

Singh, Prerna and Matthias vom Hau. 2016. "Ethnicity in Time: Politics, History, and the Relationship between Ethnic Diversity and Public Goods Provision." *Comparative Political Studies* .

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101(476):1566–1581.

Tilly, Charles. 1990. *Coercion, Capital and European States, A.D. 990-1990.* Cambridge, Mass.: Basil Blackwell.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523):1228–1242.

Wright, Joseph. 2009. "How Foreign Aid Can Foster Democratization in Authoritarian Regimes." *American Journal of Political Science* 53(3):552–571.