Online Appendix for

Characterizing and assessing temporal heterogeneity:

Introducing a change point framework, with applications on the study of democratization

Gudmund Horn Hermansen, Department of Mathematics, University of Oslo

Carl Henrik Knutsen, Department of Political Science, University of Oslo

Håvard Mokleiv Nygård, Peace Research Institute Oslo

This Appendix contains additional results that are discussed in the paper, but not reported in figures. We first present an illustration (based on a simulated example) of how conventional Chow tests may give misleading results on change point location in instances where this location is not known, but still assumed by the researcher, *a priori*. Next, we present additional results for different simulation exercises and the use of our change point method in Figures A-2 (separate change points, two different countries/parts of sample) and A-3 (gradual change over 16-year interval). Finally, we present additional results from Application II for the five regions that are not included in the discussion in the paper. Thereafter, we also include results from the extended analysis, and a discussion, on geographic heterogeneity pertaining to change points estimated separately for the time series of individual countries.

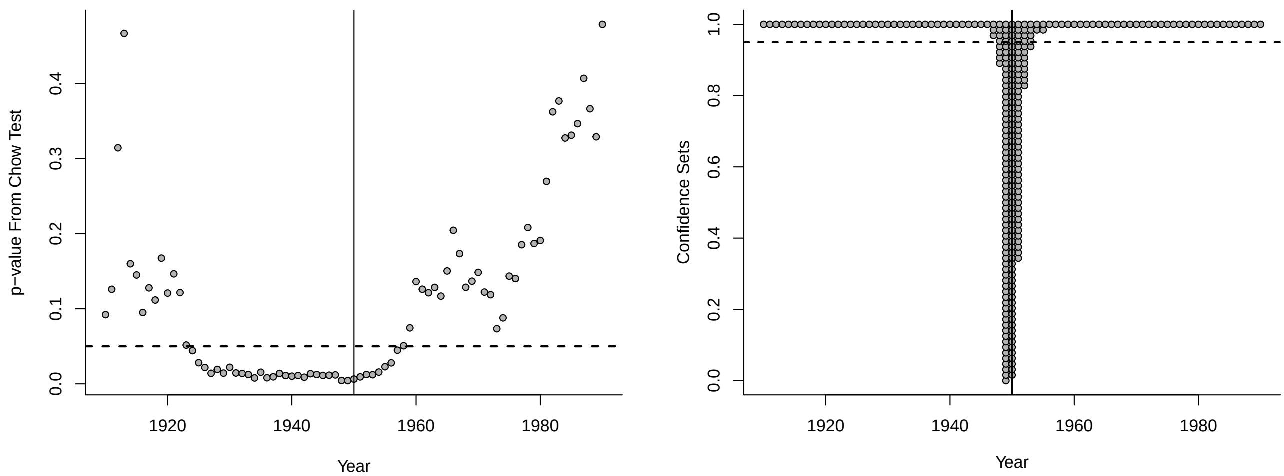
To illustrate issues with how a standard Chow test may perform in situations where the timing of the change point is not known by the researcher, *a priori*, we draw on a simulated example. More specifically, we use the same simulated data, based on the simple Model 1, that we used for Figure 1 in the paper. In this simulated example with ten fictive countries, we only have an intercept, β, to estimate. The true change-point, τ occurs in 1950, and we have that $\beta_L = 0.35$ and $\beta_R = 0.40$. The (i.i.d.) errors are set to: $\sigma = 0.10$.

In order to gauge what would happen in scenarios where the researcher does not know the true location (1950) of the change point, but can make a wide range of assumptions on when it occurred before testing for the size and statistical significance of the change point, we compute the Chow test for each year from 1910 to 1990. The corresponding p-values and a line indicating the 0.05 significance level is shown in the left panel of Figure A-1. For comparison, the corresponding confidence sets that stem from our change point method are shown in the right panel of the figure.

The smallest p-value obtained in this range is found in the same year (1949) as the point estimate generated by our procedure, which is only slightly off the true change point location. We note that running iterations across a number of possible years in this manner is a (deliberately) incorrect use of the Chow test. The Chow test is a test for a structural break at a *known* location, and is not designed to be used for estimating the *location* of a change point. Yet, in many practical instances change points are not known, but assumed by researchers based on theory, pre-existing empirical studies, or intuition. Not infrequently, such assumptions are wrong.

In the simulated example shown in Figure 1, researchers assuming that a structural break in β took place in *any particular year* in between 1922 and 1958 would have found support for this hypothesis when running a Chow test and using the conventional 0.05 significance level. When multiple tests are conducted on different assumed change points, researchers using this conventional tool would find it hard to pin down even the approximate location of the true change point. Even worse, researchers who are not conducting robustness tests on different change point locations could risk reporting strong empirical support for a structural break located far away from the true such break. In real-world examples, such patterns of results could have important substantive implications for our beliefs about the world. Incorrect theories pointing to mechanisms

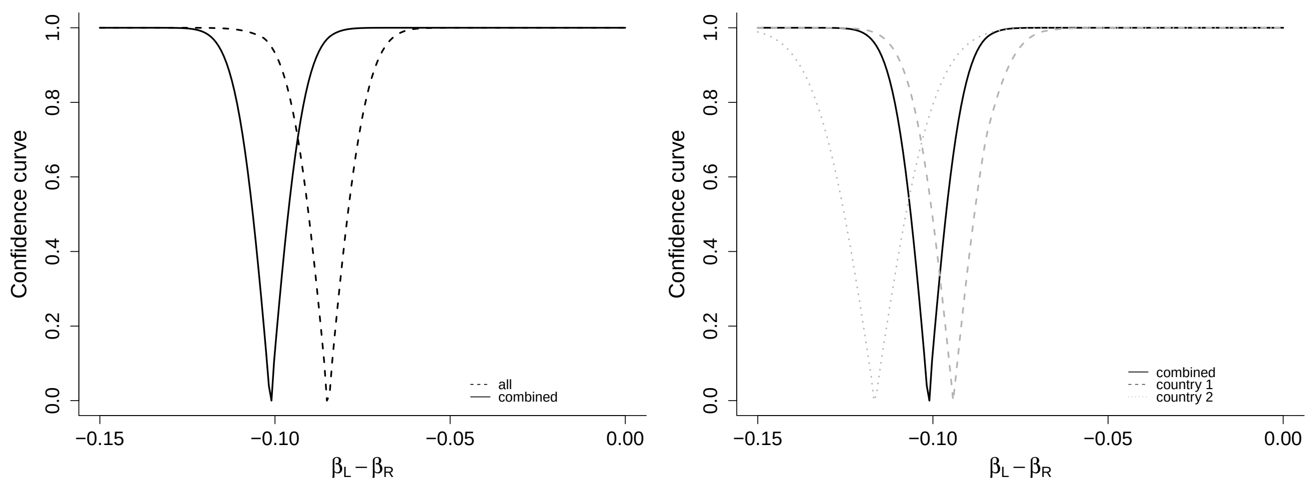
Figure A-1: Using the same simulated data from Model 1 as shown in Figure 1 we compute the Chow test for each year from 1910 to 1990. The corresponding p-values, with a dashed for $p = 0.05$, are shown in the left panel. The corresponding confidence sets from our change point method run on the same simulated data, with a dashed line indicating the 95% confidence level, are shown in the right panel.



that supposedly generate a break point during a particular time period may thus find empirical support from Chow tests, even if the theory is wrong, but there are other non-theorized mechanisms generating a change point in the same relationship, even at a very different point in time.

In contrast to the results from the iterated Chow tests, we find that only a narrow set of years (1948-1953) around the true change point are in the confidence set for the 95% confidence level when using our change point method on the same simulated data.

Figure A-2: Estimates of $\beta_L - \beta_R$. The leftmost panel displays a continuation of the 'naive' model in Figure 3, (mistakenly) assuming a joint change point for the two simulated countries as well as results for when the sample is first split into the two countries, the model is estimated, and the results are then combined. The rightmost panel takes the latter model as point of departure and shows the individual confidence curves for the two countries and the combined curve.



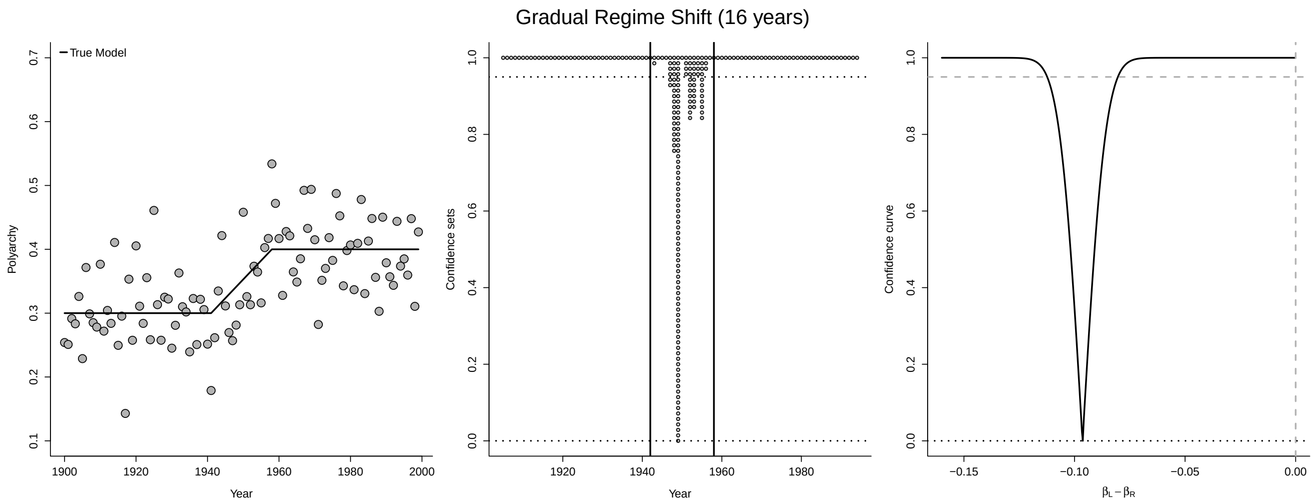


Figure A-3: Data simulated with a gradual and linear change over 16 years (from 1942 to 1958). Compared to the simulation with an 8-year change in Figure 8, the confidence sets are somewhat wider, which can be seen more clearly in the simulation study in Figure 9.

Figure A-4: Regressions on Polyarchy, with country-year as unit of analysis, sub-sampled by region, for regions not included in main paper. Does the model change over time? (Monitoring bridge, left plot). When does the relationship between GDP per capita and Polyarchy change? (Confidence sets, middle plot). What is the estimated change in the relationship? (Confidence curves for change regression coefficient; right plot).

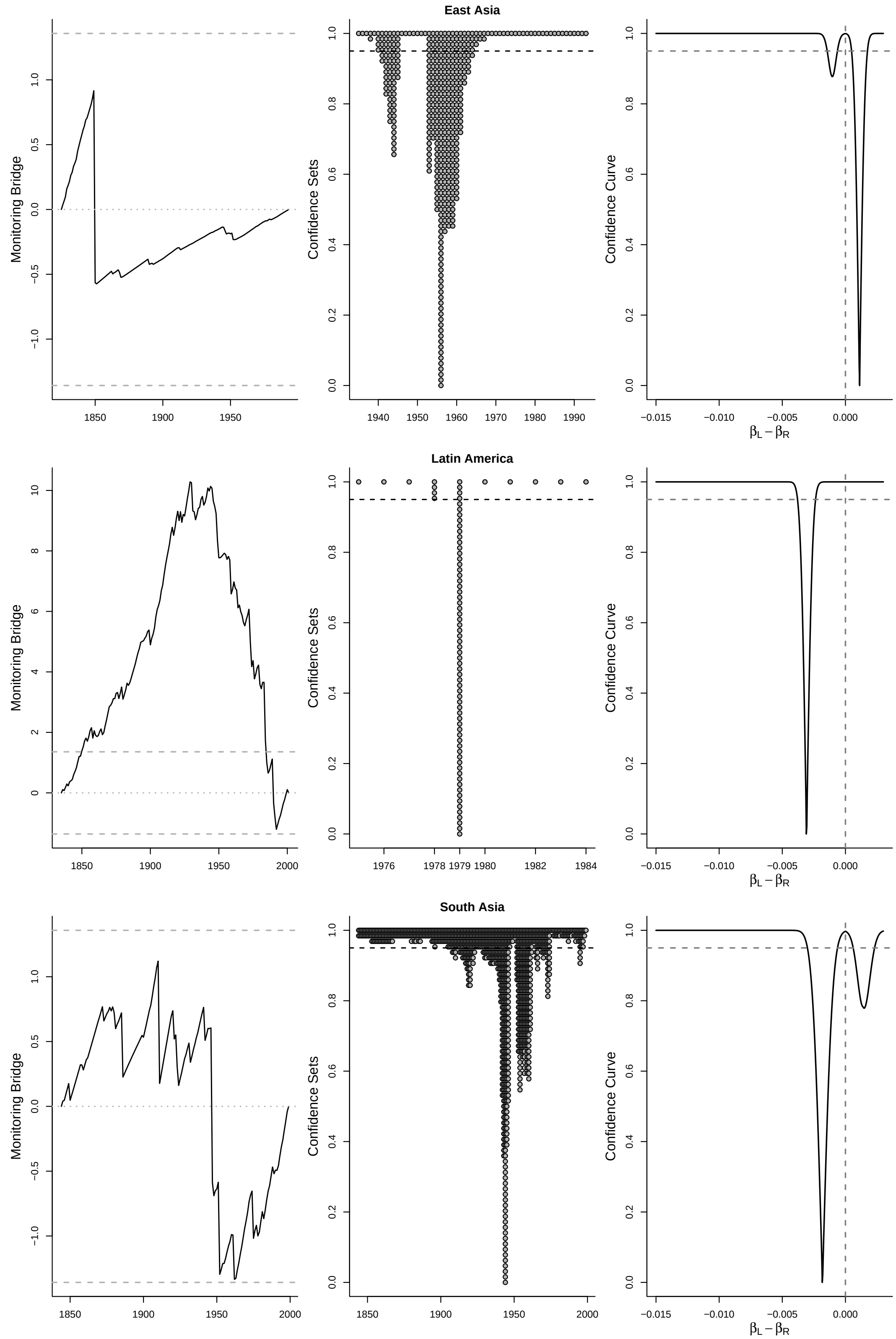
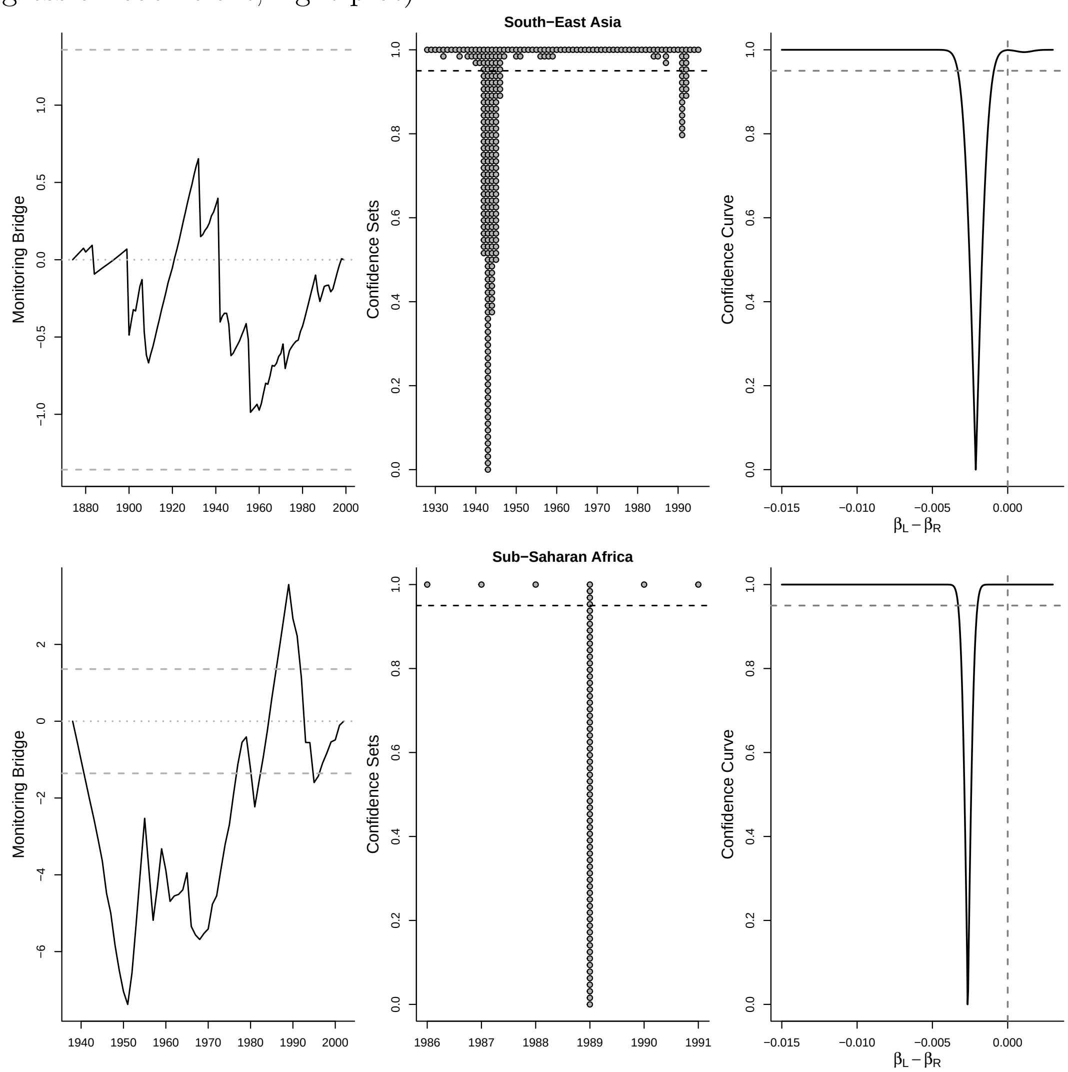


Figure A-5: Regressions on Polyarchy, with country-year as unit of analysis, sub-sampled by region, for regions not included in main paper. Does the model change over time? (Monitoring bridge, left plot). When does the relationship between GDP per capita and Polyarchy change? (Confidence sets, middle plot). What is the estimated change in the relationship? (Confidence curves for change regression coefficient; right plot).



When discussing and analyzing potential geographic/group-based heterogeneity in combination with temporal heterogeneity in our final application of the paper, we analyzed how the income–democracy relationship may differ according to (pre-specified) geographic regions of the world. Separating between regions such as Sub-Saharan Africa, Western Europe, and East Asia is standard practice for many comparative politics and international relations scholars when investigating geographic heterogeneity. Yet, it is far from given that this grouping of countries into "standard regions" is always the most plausible assumption or fruitful approach for investigating between-country heterogeneity. Countries may share other similarities than geographical proximity (language, religion, climate, topographical features, etc.) that make them susceptible to display similarities also in terms of how relationships of interest to political scientists look (and change over time). We may therefore be interested in more inductively exploring how countries (or other units, for that matter) cluster, for example in terms of timing of observed change point in a relationship, or the direction or size of the change point. Below, we therefore illustrate how one could go about in using our framework for such purposes in a very flexible manner.

In Figure A-6 we report results from our benchmark specification assessing the income–democracy relationship, separately run on sub-samples made up of observations from each individual country, for all countries with sufficiently long time series. The left panel of the Figure reports the estimated temporal location of the change point for each country, and the time interval of this change point for the 95% level of confidence. Countries are here sorted, from top to bottom, according to the estimated year of the change point, from Switzerland (before 1850) to Oman and Taiwan (close to year 2000). The right panel reports the estimated size of the change in the income–democracy relationship, and sorts countries to largest negative change (Mongolia, around -0.04) to largest positive change (Somalia, around +0.02).

Estimating separate models for each country is, in many ways, a "brute-force" tactic for handling heterogeneity, but given sufficient data it is also a very flexible tactic. For illustrative purposes, the countries in the figure are colored according to the eight regions used in the final application in the main paper. From simply eye-balling the Figure, we observe that some of the same patterns that we obtained for the regional sub-sample analysis re-appears for this more flexible

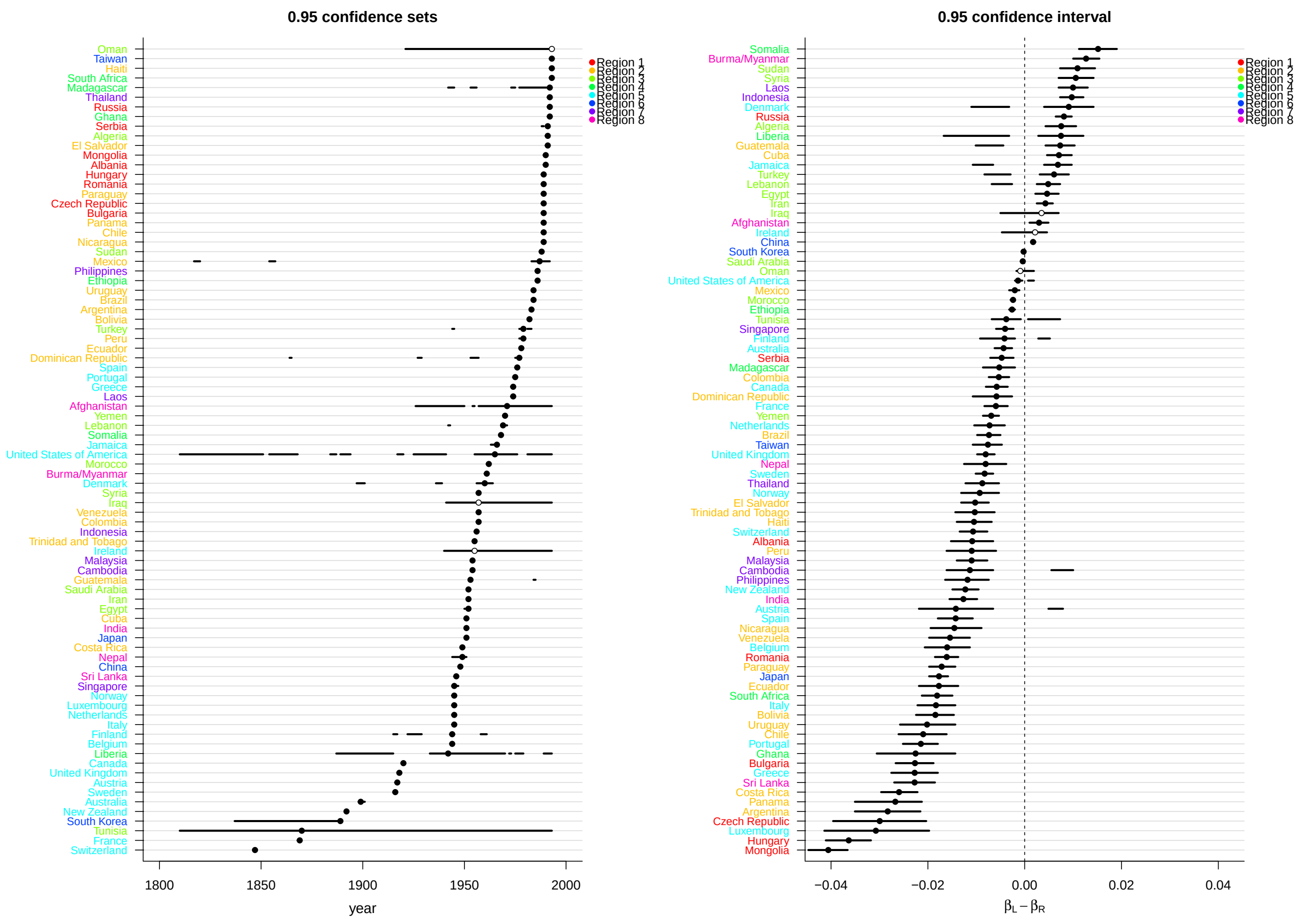


Figure A-6: Characteristics of change points for GDP per capita coefficient in regressions on Polyarchy, run separately for each country. Colors reflect country's region. Left panel displays timing of change point (with 95% confidence set). Right panel reflects estimated size of the change point (with 95% confidence interval). Please note that countries are sorted differently in the two panels.

specification. More specifically, towards the bottom of the Figure's left panel, which sorts on timing of the change point, we see a clustering of "Western" countries that all experienced relatively early change points (many of them even earlier than 1944, the regional change point detected in the paper). As we move up this panel of the figure, we see a similar "regional clustering" pattern, for Eastern European and Soviet-space countries. Many of these countries experienced a change point around 1989 (as also indicated by the region sub-sample analysis). However, between 1950 and 1989 – when most non-European countries in the sample are estimated to observe a change point – the picture is more mixed, with countries from different regions being shuffled together. Yet, we detect a clustering of Latin American countries late in the period, after the mid-1970s. Overall, however, our findings highlight that even if there are regional dynamics at play when it comes to shaping the income–democracy relationship, there remains considerable within-region heterogeneity.

The right panel reports the estimated degree of change at the country level. For the majority of countries, the estimated change is negative, as also reported in the paper for the regional and global models. This typically reflects that the relationship between income and democracy has turned stronger, positive after the change point. Nonetheless, for a handful of countries, and notably several Middle Eastern and North African ones, the estimated change in the relationship between income and democracy is actually positive, typically meaning the income coefficient was stronger positive before the change point.

Interestingly, this analysis might indicate that some countries, especially in Western Europe plus its offshoots in North America and Oceania, seem to have 'led' the rest of the world in experiencing a shift in the relationship between income and democracy. And, as noted, for the vast majority of countries, this relationship became stronger (positive) after the change point. This pattern could, for instance, reflect that the Western countries led in terms of various structural change that enhanced the significance of having high incomes for democracy, and that other countries, through a mixture of diffusion of technologies, institutions, and norms, then followed suit. This is obviously just speculation on our part, leading to (very interesting) hypotheses that need further investigation. It is well beyond the scope of this article to delve any deeper into these

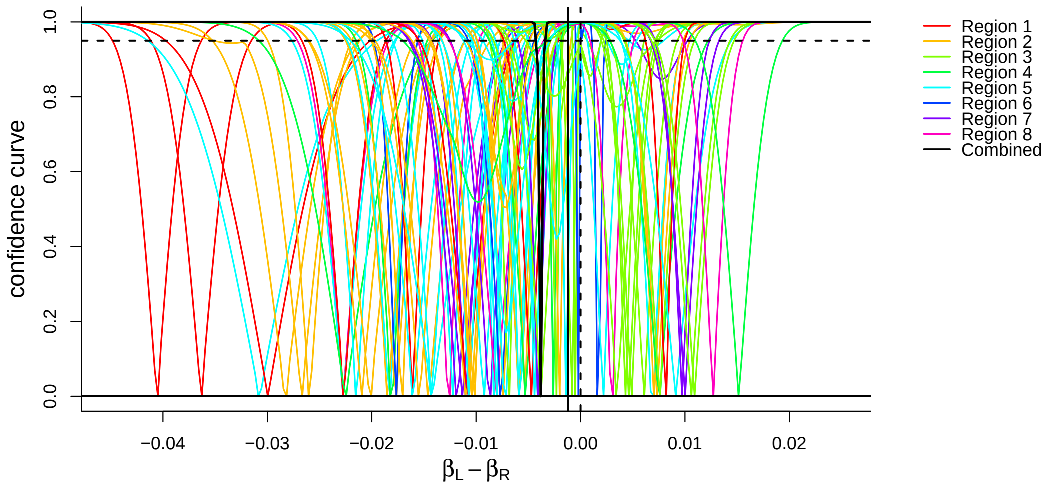


Figure A-7: Combining confidence curves for change points (for GDP per capita on Polyarchy), from separate analysis on each country. Each colored confidence curve represents one country, and colors reflect region of the country. The thicker, black confidence curve is the combined global curve, and the vertical, black line represents the estimate from the panel-specification run on the combined global sample.

substantive topics. However, this brief discussion at least illustrates how researchers can go about in assessing complex temporal and spatial dynamics and changes in relationships of interest by employing the change point framework presented here.

Finally, we show how disaggregated results may be aggregated. When running the individual time series, almost all countries the estimated degree of change is significantly different from 0. The estimate, however, varies substantially and, as discussed, even the sign of the change point is different for different countries. Using the various confidence curves we can now take all of these individual models' estimates and combine them into one (optimal) global estimate for the income coefficient. Figure thus A-7 shows the full CCs for all countries, again colored according to region, while the combined estimate is shown by the thicker, dark line. The combined estimate is negative and clearly different from 0, and its size (-0.04) is fairly close to, but still larger, negative, than what the we found for the panel-specification run on the combined global sample in the paper (-0.01; marked by a vertical black line).