Nuffield
College
UNIVERSITY OF OXFORD

ECONOMICS DISCUSSION PAPERS

Paper No: 2022-W01

Testing for Breaks in Trends with an Application to Fertility

By Jurgen A. Doornik

June 2022

Testing for Breaks in Trends with an Application to Fertility

Jurgen A. Doornik[1]*

[1] Climate Econometrics, Nuffield College, and Institute for New Economic Thinking Oxford Martin School, Oxford University, UK.

June 23, 2022

Abstract

Testing for structural breaks is an empirically relevant issue. Several procedures have been proposed in the literature, which is primarily technical. Some practical issues are addressed here, leading to a relatively simple and operational test procedure. An algorithm is proposed to find the set of breaks that minimizes the residual sum of squares in the partial structural change model, in particular a breaking trend with a fixed level. A new test is proposed that is the maximum of the supremum test of Bai & Perron (1998) and the test applied to the first differences. This controls the size for I(1) errors fairly well, and allows for the I(0) critical values to be used. Instead of supplying tables, the asymptotic and finite sample distributions are approximated. The procedures are applied to testing for breaking trends in the last 60 years of fertility data at a national and supranational level, with substantive support for broken trends in many countries.

Keywords: Asymptotic distribution; fertility, level shift; local trend; model selection; structural break.

1 Introduction

Intermittent large changes appear to be a common feature of much economic and social data. After the global financial crisis of 2008, the trend in productivity appears lower than before in many countries. Coming out of the global SARS-CoV-2 pandemic of 2020–2021, inflation has been trending upwards after staying low for a relatively long period. Testing whether a break in level or trend has occurred is of practical interest, and models covering these periods may need to accommodate several structural breaks.

One strand of literature has followed in the footsteps of Andrews (1993) by computing statistics for a break in level or trend at any point in the sample, and deriving the asymptotic distribution of the supremum of these statistics under the null of no break. Bai and Perron (1998)

*I wish to thank Jennifer L. Castle, David F. Hendry and Bent Nielsen for their valuable comments and suggestions. Financial support from the Robertson Foundation (award 9907422), Nuffield College, the Institute for New Economic Thinking (grant 20029822) and the ERC (grant 694262, DisCont) is gratefully acknowledged.

derive a test for stuctural breaks in the presence of non-breaking regressors and allowing serial correlated disturbances, not necessarily identically distributed across segments. The test can be applied to dynamic models, either with lagged dependent variables to arrive at serially uncorrelated errors, or leaving the serial correlation in the error term, but using robust standard errors. This procedure requires choosing a fraction ε of the sample size (say 10%) to impose spacing of the break point away from the sample extremes as well as other breaks. Bai and Perron (1998) provide a table of critical values based on a discrete approximation to the asymptotic distribution with a sample size of 1000 and $10\,000$ replications.

The number of breaks is not normally known, and Bai and Perron (1998) also provide a procedure for that case, first testing for no break. If that is rejected, move to the test for two breaks given one break, etc. They show that an expanding number of level shifts works, because, when the number of breaks is underspecified, the break dates are still estimated consistently. A table with critical values is supplied for this procedure for $\varepsilon = 0.05$. The assumptions underlying the derivation of the asymptotic distribution rule out trending regressors. Robust versions of the test allow for moderate amounts of stationary serial correlation.

The next stage of this literature extends the procedure of Bai and Perron simultaneously in two directions: trending behaviour with a break in the level or trend, and disturbances that can have a unit root. The initial focus is on testing for a single break at a known date, next extended to an unknown single date. Perron and Yabu (2009) use an adjusted estimate of the autoregressive parameter to construct a test statistic from feasible generalized least squares (FGLS) estimation. Any remaining serial correlation is captured by a robust estimate of the scale. Harvey, Leybourne, and Taylor (2010) provide a test for a single trend break at an unknown date, again allowing for I(0) and I(1) errors. Their test is a weighted average of robust t-tests on the broken trend of the model in levels and broken intercept in the differenced model. The weights depend on the KPSS statistic for stationarity (Kwiatkowski, Phillips, Schmidt, and Shin, 1993). This test is simpler to implement, but provides less size control in the I(1) case. A third approach is taken by Sayginsoy and Vogelsang (2011), scaling the robust test statistic by a factor that depends on the amount of serial correlation. This test requires long tables of parameters that have been calibrated by the authors.

Extensions to an unknown number of trend breaks are proposed by Kejriwal and Perron (2010) for the first procedure, and Sobreira and Nunes (2016) for the second. These apply the Bai and Perron (1998) approach: conditional on one break (say), the sample is split at this break, and each segment tested for a single break. The test is the supremum of the two. Next, conditional on two breaks, the sample is split in three etc. However, the location of the breaks is not kept fixed when a new one is added: instead, for each number of breaks, the dates are determined by the set that minimizes the residual sum of squares (RSS). This can become computationally demanding, e.g. for three breaks it is necessary to try all possible combinations of three dates subject to the adopted amount of separation ε. Bai and Perron (2003) provide a dynamic programming algorithm that is more efficient than brute force, and Kejriwal and Perron (2010) recommend using this for their procedure.

The focus of the procedures discussed so far is firmly on testing for the presence of breaks under very general conditions on the error distribution. Another approach, represented by Castle, Doornik, Hendry, and Pretis (2020), is mainly concerned with model building, with the trend (or other) breaks embedded in the model selection procedure. In the testing approach the non-breaking specification of the model is assumed known, or left unmodelled in the errors – with the latter generally the case in applications. In the modelling approach the objective is to find models with approximately normal residuals in order to sustain standard inference. We consider these complementary, and which approach is preferred depends on the empirical objectives. The asymptotic theory for the model building approach is less developed.

All the testing papers referenced sofar take a technical approach, deriving the asymptotic properties of the proposed tests. A practitioner may find it difficult to decide which procedure to use, and implementation of tests and the RSS minimization algorithm is not straightforward:

1. Asymptotic tables are based on a relatively small number of simulations and in some cases spread over several publications, partially in unpublished papers, or embedded in downloadable code. The choice of sample truncation ε may also be limited to just one value.

2. Each procedure uses a different method to compute robust estimates, and it is unclear to what extent this affects the results in realistic samples. All three basic procedures use different kernels in the robust estimates.

3. The algorithm of Bai and Perron (2003) does not apply to the model where only the trend breaks. From an empirical perspective this may be the more relevant model, so arguably no practical solution has been proposed yet in the literature.

4. Even with a working algorithm there remains the issue that, when the number of broken trends is underspecified, trend dating is not always consistent – unlike the model with level shifts. This is illustrated by Yang (2017), who argues that break point estimation in the first differenced model can be used instead. Although, provided the test still rejects with a wrongly-dated break, the overall procedure can have power because it does not condition on this wrong date.

We present an alternative algorithm to estimate the break dates that is valid in a more general setting in §2. It can also be substantially faster. Then §3 studies the new algorithm in simulation experiments with two breaks in the DGP. A modified test is proposed in §4, which also works for I(1) errors. Critical values are provided through simple approximations to the asymptotic distributions, similar to Nielsen (1997) and Doornik (1998). §5 studies the new test and distributional approximations through simulations; §6 considers testing for multiple breaks. The empirical application in §7 tests for breaks in fertility and birth rates in most countries of the world. §8 concludes, while some further results are in appendices.

2 An algorithm to estimate break dates

We show that the algorithm of Bai and Perron (2003) (BP03) cannot be used when only the trend breaks, and introduce an alternative algorithm.

The basic data generation process is one with a break in period B_1 and an autoregressive error term:

$$y_t = \boldsymbol{d}_t'\boldsymbol{\beta} + \boldsymbol{x}_t(B_1)'\boldsymbol{\psi}(B_1) + u_t, \quad u_t = \rho u_{t-1} + \epsilon_t, \tag{1}$$

for $t = 1, ..., T$, with u_0 fixed, ϵ_t stationary white noise, and in most cases $|\rho| \leq 1$, so I(1) errors are allowed. The regressors $\boldsymbol{d}_t$ and $\boldsymbol{x}_t(B_1)$ contain q and k known deterministic terms respectively, and we are interested in testing H_0: $\boldsymbol{\psi}(B_1) = \mathbf{0}$. In matrix form: $\boldsymbol{D}' = (d_1, ..., d_T)$ and $\boldsymbol{X}(s)' = (x_1(s), ..., x_T(s))$. Regardless of the structure of u_t, the first step is ordinary least squares (OLS) estimation of

$$y_t = \boldsymbol{d}_t'\boldsymbol{\beta} + \boldsymbol{x}_t(B_1)'\boldsymbol{\psi}(B_1) + u_t, \quad t = 1, ..., T. \tag{2}$$

Particular specifications with a break at s allow for broken intercept $U_t(s) = 1_{\{t>s\}}$ and broken trend $D_t(s) = (t - s)U_t(s)$, numbered here as follows:

$$\begin{array}{llll} \text{Model 0:} & \boldsymbol{d}_t = 1 \quad \boldsymbol{x}_t(s) = U_t(s) & \text{(broken intercept, no trend),} \\ \text{Model 1:} & \boldsymbol{d}_t' = (1, t) \quad \boldsymbol{x}_t(s) = U_t(s) & \text{(broken intercept),} \\ \text{Model 2:} & \boldsymbol{d}_t' = (1, t) \quad \boldsymbol{x}_t(s) = D_t(s) & \text{(broken trend),} \\ \text{Model 3:} & \boldsymbol{d}_t' = (1, t) \quad \boldsymbol{x}_t(s)' = (U_t(s), D_t(s)) & \text{(both broken).} \end{array} \tag{3}$$

There may be additional non-breaking regressors, which are added to $\boldsymbol{d}_t$. Model 0 is handled in Bai and Perron (1998) (BP), where trending regressors and $\rho = 1$ are ruled out from the test.

An important aspect is the estimation of break dates, in this case minimizing the RSS for the given number of breaks. The dynamic programming algorithm of BP03 applies to the pure structural change model, i.e. all regressors break, which can be written with block-partitioned regressors. Extending (2) to m breaks:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{B}}^{\#}\boldsymbol{\theta} + \boldsymbol{u}_{\mathcal{B}}. \tag{4}$$

Here $\boldsymbol{X}_{\mathcal{B}}^{\#} = \text{diag}(\boldsymbol{X}_1, ..., \boldsymbol{X}_{m+1})$ for break set $\mathcal{B}(m) = (B_1, ..., B_m)$, $B_i \leq B_{i+1} - T_0$, and $k(m+1)$-column vector $\boldsymbol{\theta}$, $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_{m+1}')$. The algorithm finds the optimal set of m breaks by OLS in terms of RSS:

$$\widehat{\mathcal{B}}_{\varepsilon}(m) = \underset{B_1,...,B_m \in [T_0, T_1]}{\arg\min} \left\{ \widehat{\sigma}_u^2(\mathcal{B}) = \tfrac{1}{T}\boldsymbol{u}_{\mathcal{B}}'\boldsymbol{u}_{\mathcal{B}} \mid \boldsymbol{y} = \boldsymbol{X}_{\mathcal{B}}^{\#}\boldsymbol{\theta} + \boldsymbol{u}_{\mathcal{B}}; \varepsilon \right\}. \tag{5}$$

The cut-off ε defines the sample span $\lfloor \varepsilon T \rfloor, ..., \lfloor (1 - \varepsilon)T \rfloor$ in which the breaks may fall, while the breaks must also be $\lfloor \varepsilon T \rfloor$ observations apart, where $\lfloor \rfloor$ truncates down to the nearest integer.

Models 0 and 3, see (3), clearly fit the pattern of (4), with the null hypothesis of no break obtained by $\boldsymbol{\theta}_1 = ... = \boldsymbol{\theta}_{m+1}'$. Implementation of the BP03 algorithm requires storage of order T^2, and $T - T_0$ recursive least squares (RLS) runs. Recursive computation of RSS can be implemented efficiently by downdating the QR decomposition (see Appendix A), which makes the algorithm relatively fast.

For the partial structural change model, i.e. models 1 and 2, the non-breaking variables are captured in the matrix $\boldsymbol{D}$.

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{B}}^{\#}\boldsymbol{\theta} + \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{u}_{\mathcal{B}}, \tag{6}$$

BP03 suggest an iterative application of the dynamic programming algorithm, alternating between $\boldsymbol{\theta}, \mathcal{B}|\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}, \boldsymbol{\theta}|\widehat{\mathcal{B}}$. This allows for the dynamic programming algorithm to be used on the pure conditional change model:

$$\boldsymbol{y} - \boldsymbol{D}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}_{\mathcal{B}}^{\#}\boldsymbol{\theta} + \boldsymbol{u}_{\mathcal{B}}. \tag{7}$$

Then $\boldsymbol{\beta}$ is updated with (6) based on new break set (as is $\boldsymbol{\theta}$, but its value is not needed).[1] We refer to this procedure as the BP03 algorithm. Convergence to a stationary point is achieved when no further progress is made in the alternating variables iteration. This approach can be used for model 1, because the broken intercept can be block partitioned.

However, this iterated dynamic programming solution for (7) does not apply to model 2, because the broken trend $D_t(s) = (t - s)1_{\{t>s\}}$ starts at one when $t = s + 1$. Consider, e.g., a single break half-way in a sample of six, which has regressors (ignoring the intercept):

$$\text{model 2:} \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 4 & 1 \\ 5 & 2 \\ 6 & 3 \end{pmatrix} \qquad \text{BP03 algorithm:} \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 0 & 4 \\ 0 & 5 \\ 0 & 6 \end{pmatrix}.$$

So the BP03 algorithm for a trend break only, which has intercept and trend in $\boldsymbol{D}$, also breaks the intercept in a specific way. This does not correspond to model 2.

Appendix A proposes an alternative algorithm that applies to all models, including model 2. In each iteration it removes a current break point, one at a time, to see if it can be replaced by a better one. This algorithm converges when the break set does not change, which is a stationary point. A global maximum is not guaranteed. Special care is required to enforce separation of the break points.

As an illustration we use the ex-post interest rate of Garcia and Perron (1996). We implemented the BP03 algorithm in Ox 9 (Doornik, 2021), and have verified the results with the GAUSS code available from Pierre Perron's web site at `blogs.bu.edu/perron/codes`. This code uses the same data set, and allows any variables to be specified as breaking or non-breaking. The dashed line in the top-left plot in Figure 1 shows the estimated trend with a single break from the BP03 algorithm. Our new algorithm is shown in the solid line. The top-right plot repeats this for two breaks. In both cases we see that the application of BP03 does not give the intended result.

The bottom plots in Figure 1 offer the estimated breaks using the model 3 specification in comparison. In this case, both algorithms are in complete agreement.

We should emphasize that neither algorithm guarantees a global minimum. For example, the BP03 algorithm applying model 1 to the ex post real interest rate spaced by ten observations finds three breaks at $\mathcal{B} = (24, 47, 79)$ with RSS 443.1. However, break set $(47, 57, 79)$ has a

[1] BP03 §3.5 suggests to get the initial values of $\mathcal{B}$ and $\boldsymbol{\theta}$ by treating all variables as breaking:

$$\boldsymbol{y} = \boldsymbol{X}^{\#}_{\mathcal{B}}\boldsymbol{\theta} + \boldsymbol{D}^{\#}_{\mathcal{B}}\boldsymbol{B} + \boldsymbol{v}_{\mathcal{B}},$$

then estimating $\boldsymbol{\beta}$ from:

$$\boldsymbol{y} - \boldsymbol{X}^{\#}_{\widehat{\mathcal{B}}}\widehat{\boldsymbol{\theta}} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{w}.$$

This can result in poor starting values, e.g. if we happen to start with the correct break set, it will move the initial $\boldsymbol{\beta}$ far from the optimal value. An alternative is to use only the initial break set, and get $\boldsymbol{\beta}$ from (7). To illustrate the difference, we use Model 1 for the ex-post real interest rate and spacing of $\lfloor \varepsilon T \rfloor = \lfloor 0.05 \times 103 \rfloor$. The starting RSS for three breaks is 57208.3 using BP03 §3.5. But with our suggestion the initial RSS is 491.0, saving one iteration.

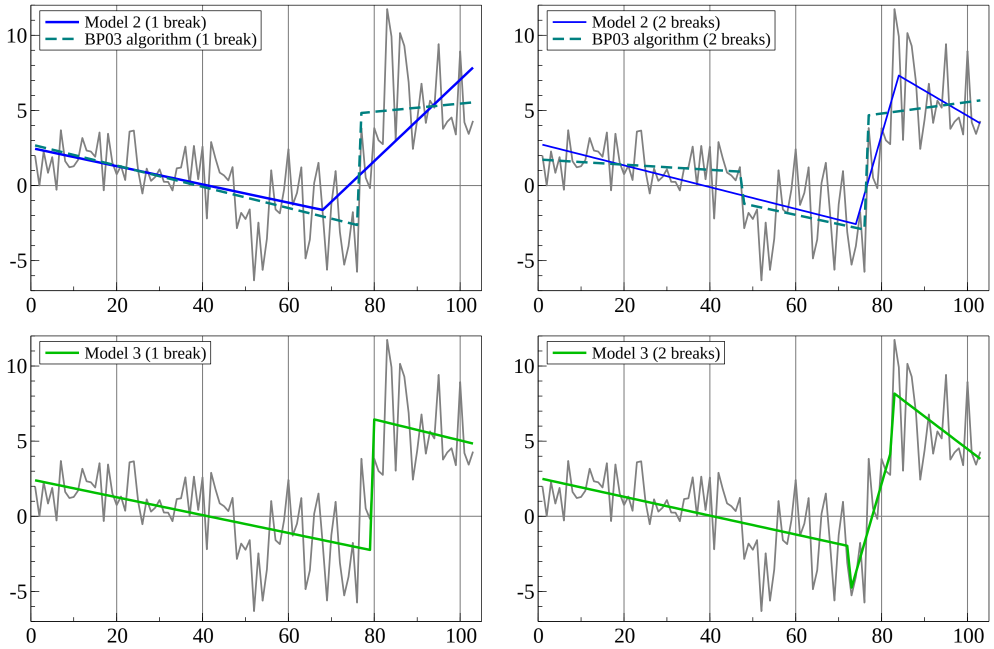

Figure 1: BP data. Optimal break set for model 2 (top) and model 3 (bottom) with one break (left) and two breaks (right), together with optimal break dates when the BP03 algorithm is applied to model 2.

lower RSS of 436.0. This is more likely to happen when the number of breaks is over specified, and then less relevant empirically.

3 Simulations for break dating algorithms

The performance of the dating algorithms is studied in a simulation experiment. The results are presented in graphs.

To compare the BP03 algorithm to the proposed algorithm, we adopt a DGP similar to Yang (2017) This has two breaks in the DGP, while searching for up to three break dates.

First define the DGP used in Perron and Yabu (2009) with up to one break:

$$\begin{aligned} y_t &= \eta_1 U_t(s) + \eta_2 D_t(s) + u_t, && t = 1, ..., 100, \\ u_t &= \rho u_{t-1} + \zeta \Delta u_{t-1} + \epsilon_t + \theta \epsilon_{t-1}, && u_{-1} = u_0 = \epsilon_0 = 0, \epsilon_t \sim N[0,1]. \end{aligned} \tag{8}$$

The breaks are half-way so $s = T/2$, and under the null of no break: $\eta_1 = \eta_2 = 0$. Extending (8) to M^* breaks:

$$y_t = \sum_{m=1}^{M^*} \eta_{1,m} U_t(\lfloor mT/M^* \rfloor) + \sum_{m=1}^{M^*} \eta_{2,m} D_t(\lfloor mT/M^* \rfloor) + u_t. \tag{9}$$

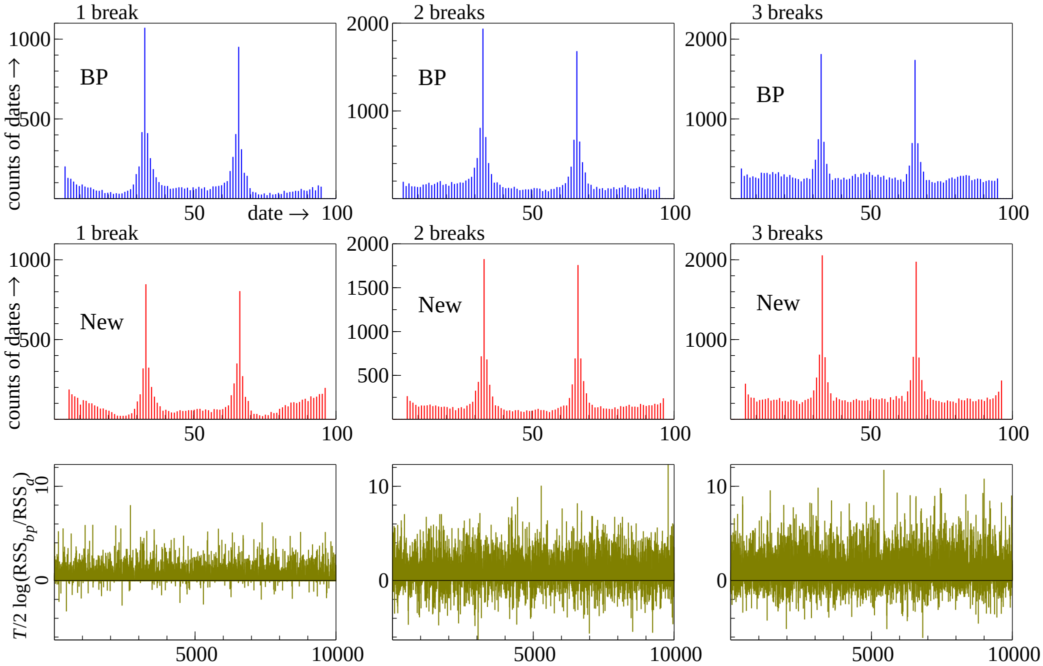

Figure 2: Break dates for Model 1 with two level shifts in the DGP (10). Count of detected dates with BP03 algorithm at the top, counts with alternative algorithm in the middle, difference in log-likelihoods at the bottom.

Simulations in this section use independent normal errors ($\rho = \zeta = \theta = 0$), and two breaks at observations 33 and 66 for a sample size of 100, using the following parameter values:

$$\text{DGP1 for Model 1:}\quad \eta_{1,1} = \eta_{1,2} = 1, \eta_{2,1} = \eta_{2,2} = 0, \tag{10}$$

$$\text{DGP2 for Model 2:}\quad \eta_{1,1} = \eta_{1,2} = 0, \eta_{2,1} = \eta_{2,2} = 1, \tag{11}$$

$$\text{DGP3 for Model 3:}\quad \eta_{1,1} = \eta_{1,2} = 1, \eta_{2,1} = \eta_{2,2} = 1. \tag{12}$$

We set $\varepsilon = 0.05$.

The results are in Figures 2–4. The first column of graphs in each figure has the results when searching for one break, the middle column hs the results for two breaks (which corresponds to the DGP) and the final column is for three breaks. The top row graphs shows the count of dates found using the BP03 algorithm in 10 000 replications. The middle row reports this for the proposed algorithm.

There is not much difference between the detected dates of the two algorithms in Figure 2. The breaks are detected even when the number of breaks is underspecified (in the presence of a trend, but also without), which improves as the sample size grows. This consistent selection in the underspecified case is derived theoretically in BP.

The bottom row of Figure 2 compares the log-likelihood of the two methods for each repli-

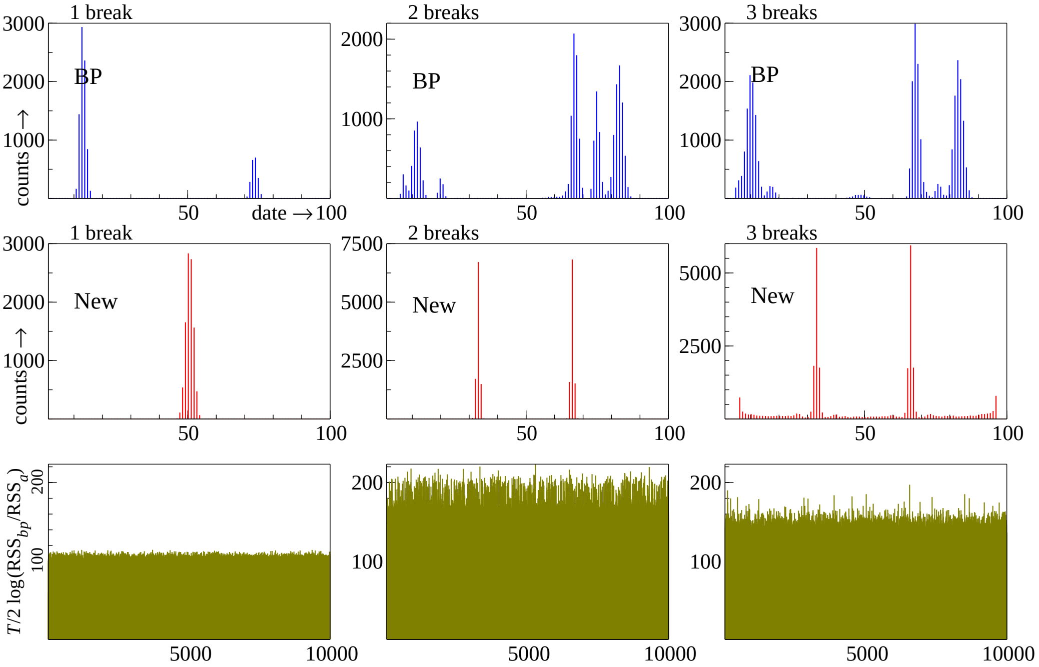

Figure 3: Break dates for Model 2 with two trend shifts in the DGP (11). BP03 algorithm at the top, alternative algorithm in the middle, difference in log-likelihoods at the bottom.

cation (using the same generated data in each case):

$$\frac{T}{2} \log \frac{\mathrm{RSS}_{bp}}{\mathrm{RSS}_{a}}.$$

If this is positive, then we prefer the alternative algorithm, otherwise we prefer the BP03 algorithm.

In model 2 only the trend breaks. The top row of Figure 3 reconfirms that the BP03 algorithm is invalid here. The new algorithm performs well, but it shows what Yang (2017) discusses: trend break selection in the underspecified case is not consistent. This also applies to model 3, which is a pure structural change model where the BP03 algorithm is valid, see Figure 4.

The new algorithm completes the simulation in about half the time of the BP03 algorithm for the pure structural change model. For the partial model the new algorithm is about 15 to 20 times faster (this will be less for a larger ε).

Yang (2017) proposes to date a broken trend in the differenced model. For model 2 this also means that the BP03 algorithm could be used. The top row of Figure 5 shows that the underspecified model is consistent again in that case. The first difference of model 3 introduces a moving impulse dummy, which can be handled in the new algorithm (middle row). However, the bottom row of Figure 5 shows that it is better to just ignore this impulse dummy when it comes to dating, at least in this setting.

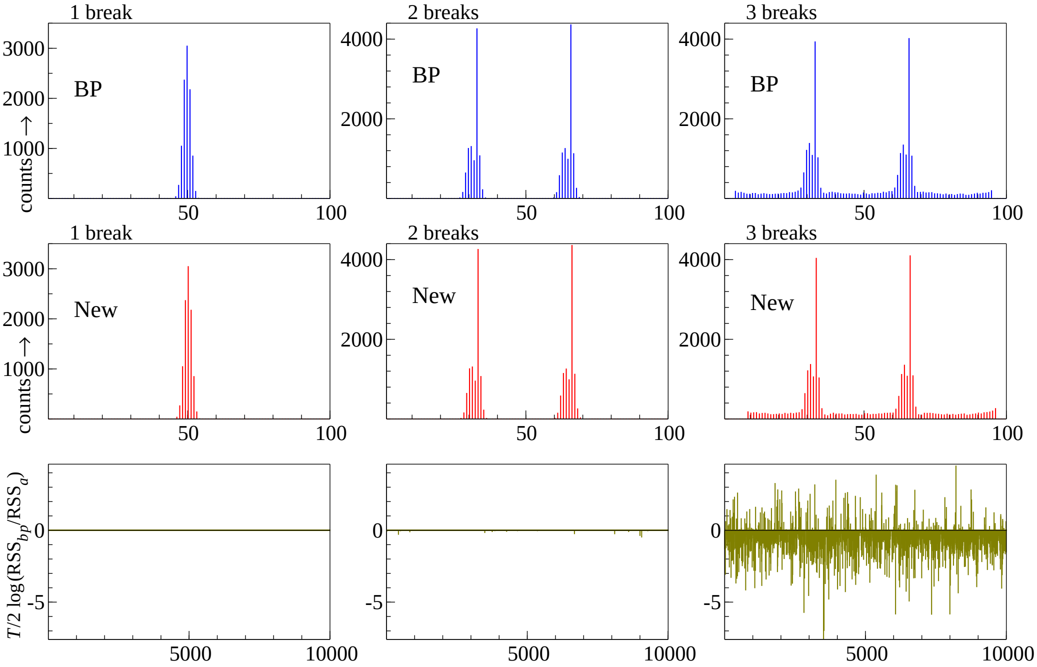

Figure 4: Break dates for Model 3 with two trend and level shifts in the DGP (12). BP03 algorithm at the top, alternative algorithm in the middle, difference in log-likelihoods at the bottom.

4 Tests for a single break at an unknown date

A new test is proposed that is the maximum of the supremum test of Bai & Perron and the test applied to the first differences. This offers some control over the size for I(1) errors. Approximations to the distributions are obtained, which give critical values, as well as p-values.

The Bai and Perron (1998) test is based on the Wald test for $\boldsymbol{\psi}(s) = \mathbf{0}$ at observation s in (2), estimated by OLS:

$$\begin{aligned} W(s) &= \widehat{\boldsymbol{\psi}}'\mathrm{V}\Big[\widehat{\boldsymbol{\psi}}\Big]^{-1}\widehat{\boldsymbol{\psi}}\,\frac{T-k-q}{T} \\ &= \widehat{\sigma}^{-2}\widehat{\boldsymbol{\psi}}'\left[\boldsymbol{X}(s)'\boldsymbol{X}(s) - \boldsymbol{X}(s)'\boldsymbol{D}\left(\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{D}'\boldsymbol{X}(s)\right]\widehat{\boldsymbol{\psi}}\,\frac{T-k-q}{T}, \end{aligned} \tag{13}$$

where $\widehat{\sigma}_u^2 = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}/T$, and k, q are the number of columns in $\boldsymbol{X}(s)$ and $\boldsymbol{D}$ respectively. The scale factor amounts to replacing $\widehat{\sigma}_u^2$ by the OLS version $\widetilde{\sigma}_u^2 = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}/(T-k-q)$.

The best break date $\widehat{s}$ is obtained by minimizing[2] the RSS from (2). This leads to the basic version of the BP supremum test, labelled BPN here:

$$\mathrm{sup}F_{\mathrm{BPN}} = W(\widehat{s}). \tag{14}$$

[2] It is also possible to compute the test for all possible breaks. However, BP suggests this procedure is asymptotically equivalent. It is also more convenient and faster.

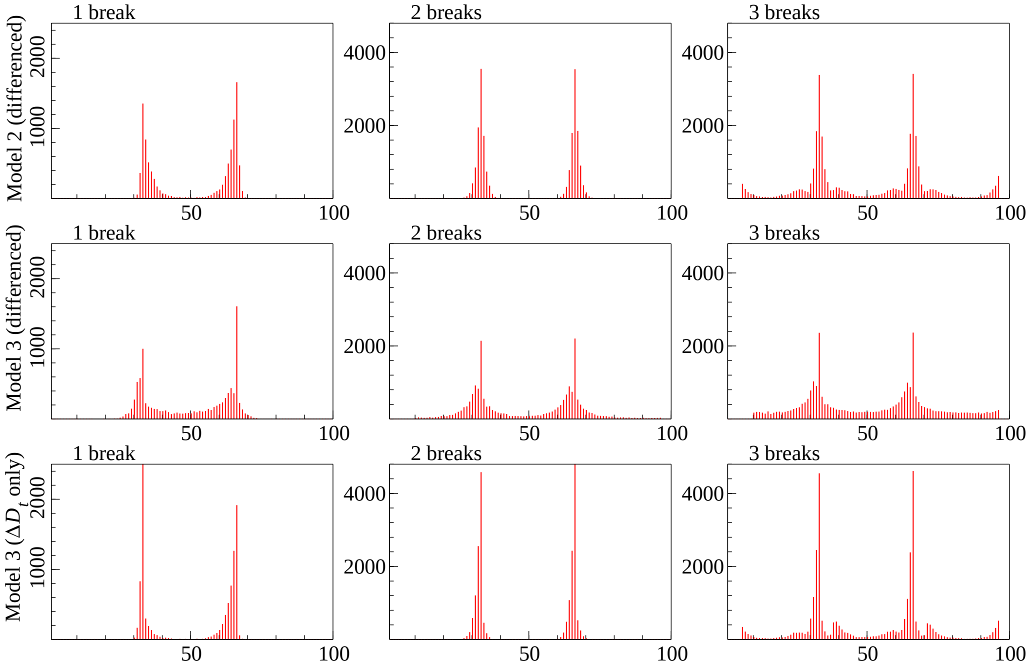

Figure 5: Break dates for Model 2 and 3 in differences (top and middle), and model 3 in differences without the impulse dummy. DGP (11) and (12).

BP proposes several extensions to serially correlated errors, as well as heteroscedasticity, which we ignore here. One suggestion is to construct a robust version of the entire $\mathrm{V}\left[\widehat{\psi}\right]$. However, that breaks down with trending regressors. We only consider robust estimation of the scale through the procedure of Andrews (1991) with prewhitening. The prewhitening is a simple first order autoregression, AR(1), for $\widehat{u}_t$, estimated by OLS to give $\widehat{\rho}$, truncated to $|\widehat{\rho}| \leq 1-1/T \equiv \bar{\rho}$. Then we apply the quadratic spectral kernel estimator with plug-in automatic bandwidth and the same truncation. The kernel length is truncated at twenty times the bandwidth; often 50 is used, but we truncate harder because this is faster without noticeable impact. Write $\widehat{h}_u$ for the robust scale so computed, and $\widehat{h}_u^0$ if prewhitening is omitted. The BP test using the quadratic spectral kernel for the scale is denoted BPQ:

$$\sup F_{\text{BPQ}} = \sup F_{\text{BPN}} \frac{\widehat{\sigma}_u^2}{\widehat{h}_u}. \tag{15}$$

Subsequently, we shall apply the test with an underspecified number of breaks. This can invalidate robust estimation, as the neglected partial trends may be interpreted as serial correlation by the robust procedure even when the errors are white noise. This may also affect the power in the procedure of Sobreira and Nunes (2016), because omitted partial trends can lead the KPSS statistics to put all weight on the models in differences.

While $\sup F_{\text{BPQ}}$ improves over $\sup F_{\text{BPN}}$ for moderate amounts of serial correlation, its size is close to one if there is a unit root. In the introduction we refer to three proposals that handle this in different ways. Perron and Yabu (2009) construct a Wald test for each possible observation. First estimate ρ from the first stage OLS residuals, apply a bias correction that shifts the estimates towards the unit root in small samples, and then set this to one if the bias corrected $\widehat{\rho}$ exceeds $1 - T^{-1/2}$. The Wald statistic is computed in the FGLS model. Finally, a choice between three robust estimates of scale is made, depending on the estimated dynamics of the residuals. They consider both the supremum version and the geometric average, preferring the latter because the I(1) and I(0) distributions are close. With the lag length selection and robust estimation computed at each possible break date, the test takes more time.

We propose a simpler test, which is an extension of the Bai & Perron test. While it is a supremum version, it also means that the original distribution can still be applied in the I(0) case. This is achieved by adding in the test based on the differenced model, but without any parameter driven weighting.

We follow Perron and Yabu (2009) in adopting the bias correction procedure of Roy and Fuller (2001), but with fixed parameters. Let $\widehat{\rho}$ be the OLS estimate using residuals $\widehat{u}_t$ from (2):

$$\widehat{u}_t = \rho \widehat{u}_{t-1} + v_t, t = 2, ..., T. \tag{16}$$

Use $\widehat{\rho}$, restricted to $[-0.99, 1]$, and its estimated variance $\widehat{\sigma}^2_\rho$ to construct the t-ratio $\widehat{t} = (\widehat{\rho} - 1)/\widehat{\sigma}_\rho$. The bias corrected version is $\widehat{\rho}_c$, defined as:

$$\widehat{\rho}_c = \widehat{\rho} + C\widehat{\sigma}_\rho,$$
$$C = \begin{cases} -\widehat{t} & \text{if } \widehat{t} > \tau_1, \\ t^* - K/[\widehat{t} + c_2(\widehat{t} - \tau_2)] & \text{if } \tau_2 < \widehat{t} \leq \tau_1, \\ t^* - K/\widehat{t} & \text{if } \tau_3 < \widehat{t} \leq \tau_2, \\ 0 & \text{if } \widehat{t} \leq \tau_3, \end{cases}$$

where $n = \lfloor (p+1)/2 \rfloor$, $t^* = n\widehat{t}/T$, $K = q+k+1$, $c_2 = [KT - \tau_1^2(n+T)]/[\tau_1(\tau_1 - \tau_2)(n+T)]$, $\tau_1 = -4$, $\tau_2 = -10$, $\tau_3 = -(KT)^{1/2}$. Finally, p is the lag length used in (16), but we set it to zero. The value of τ_1 in Perron and Yabu (2009) varies between -4.9 and -4.2 with the relative location of $\widehat{s}$ in the sample and the model; here it is kept fixed.

The new test procedure is as follows. Estimate (2) to obtain $W(\widehat{s})$ and $\widehat{u}_t$, then estimate (16) and construct $\widehat{\rho}_c$ whitened residuals:

$$\widehat{e}_t = \widehat{v}_t - \widehat{\rho}_c \widehat{v}_{t-1}, t = 2, ..., T.$$

The first robust statistic is:

$$W_1(\widehat{s}) = \begin{cases} W(\widehat{s})\,(1 - \widehat{\rho}_c)^2\ \widehat{\sigma}_u/\widehat{h}^0_e, & \text{if } \widehat{\rho}_c < \bar{\rho}, \\ W(\widehat{s})\,(1 - \bar{\rho})^2\ \widehat{\sigma}_u/\widehat{h}^0_e, & \text{otherwise.} \end{cases}$$

The bias corrected value is used for prewhitening the robust scale estimate, and h^0 is the quadratic spectral estimate without further prewhitening; $\bar{\rho} = 1 - 1/T$. This improves the properties close to a unit root, where the method employed in supF_{BPQ} is less effective.

The Wald test is also computed for the model in differences:

$$\Delta y_t = \Delta \boldsymbol{d}_t' \boldsymbol{\beta}_\Delta + \Delta \boldsymbol{x}_t(s)' \boldsymbol{\psi}_\Delta(s) + w_t, \quad t = 2, ..., T, \tag{17}$$

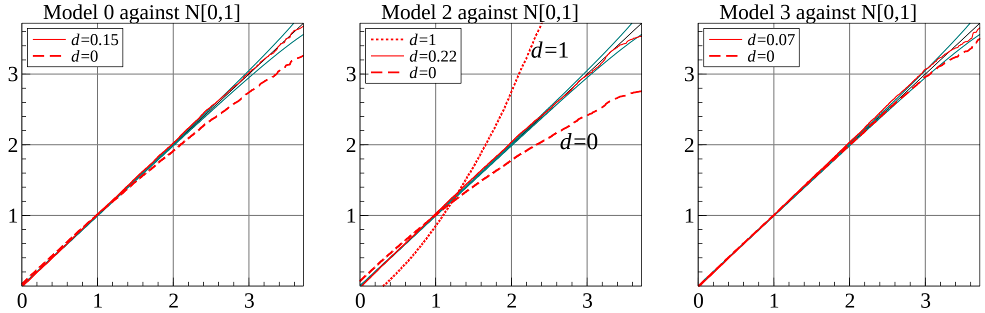

Figure 6: QQ plots of transformations of the test statistic supF_{MAX}. Model 1 (left), Model 2 (middle), Model 3 (right), all against right-half of N$[0,1]$. QQ plots truncated at quantiles 0.9999. $\varepsilon = 0.1$.

giving

$$W_\Delta(s) = \widehat{\psi}'_\Delta \mathrm{V}\left[\widehat{\psi_\Delta}\right]^{-1} \widehat{\psi}_\Delta \, \frac{T-k-q}{T-1},$$

and additional statistic:

$$W_2(\widehat{s}) = \begin{cases} W_\Delta(\widehat{s}) & \text{if } \widehat{\rho}_c < \bar{\rho}, \\ W_\Delta(\widehat{s})\dfrac{\widehat{\sigma}_w}{h_{\widehat{w}}} & \text{otherwise.} \end{cases}$$

The new statistic takes the maximum of the two supremum tests:

$$\mathrm{sup}F_{\mathrm{MAX}} = \max\left\{W_1(\widehat{s}), W_2(\widehat{s})\right\}. \tag{18}$$

When the model is stationary and not close to a unit root, the test in levels dominates, and the same distribution applies as when using just the supremum test. When the model has a unit root, or is close, the test in differences will dominate. This changes the trend and broken trend to intercept and broken intercept, so the test for that model will dominate. This is close enough to use the distribution as if we used just the model in levels, and also keeps the distribution reasonably close for a unit root. Avoiding estimating and calibrating a weighting function simplifies the procedure, while we do not need to change the critical values.

Instead of tabulating we approximate the asymptotic distribution through a response surface. We start by simulating $100\,000$ values of supF_{MAX} at sample size $T = 1000$ under i.i.d. normality. QQ plots of power transformations of the recorded test outcomes are inspected to get a decent match to normality in the right tail of the distribution. This gives power $d = 0.15$ for model 0, $d = 0.12$ for model 1, and $d = 0.22$ for model 2. Model 3 is close to logarithmic (denoted by power $d = 0$) with $d = 0.07$. Figure 6 shows some of the QQ plots. On the left are the power-transformed tests under model 1 using two different values, then standardized. The middle plot is for model 2, using three values for d. In that case $d = 0.22$ works well in the right tail. (Remember that the Wilson-Hilferty transformation of a χ^2 random variable to approximate normality has $d = 1/3$.) Point-wise confidence bands are based on Engler and Nielsen (2009).

Next, we simulate the means and standard deviations from the transformed test at sample sizes $60, 100, 125, 150, 200, 250, 500, 1000$ for ε in $\{0.01, 0.03, 0.05, 0.07, 0.10, 0.15, 0.20\}$ using $10\,000$ replications, provided $\lfloor \varepsilon T \rfloor \geq 5$. Finally, a response surface is fitted to these 44

			Coefficients for m and s equations				
Model	d	m, s	$100/T$	$(100/T)^2$	ε	$\varepsilon^{1/2}$	1
0	0.15	$m(T,\varepsilon) =$	-0.00494	0.00326	-0.0413	-0.152	1.30
0	0.15	$s(T,\varepsilon) =$	0.0103	-0.00104	0.00804	0.0457	0.0814
1	0.12	$m(T,\varepsilon) =$	-0.00722	0.00494	-0.0407	-0.0597	1.25
1	0.12	$s(T,\varepsilon) =$	0.00880	-0.000628	-0.0104	0.0281	0.0570
2	0.22	$m(T,\varepsilon) =$	0.0270	-0.00327	-0.0299	-0.472	1.31
2	0.22	$s(T,\varepsilon) =$	0.0101	-0.00126	-0.0210	0.143	0.165
3	0.07	$m(T,\varepsilon) =$	-0.00198	0.00264	-0.0539	-0.0359	1.17
3	0.07	$s(T,\varepsilon) =$	0.00547	-0.000674	0.0139	0.0118	0.0250

Table 1: Parameters m, a for the approximation to the mean and standard deviation of the transformed statistic.

'observations'. The p-values are computed from the approximating standard normal cdf:

$$p_{\text{app}}(T,\varepsilon;d) = \begin{cases} 1-\Phi\left(\frac{[\sup F_{\text{MAX}}]^d - m(T,\varepsilon)}{s(T,\varepsilon)}\right) & \text{if } d>0, \\ 1-\Phi\left(\frac{\log[\sup F_{\text{MAX}}] - m(T,\varepsilon)}{s(T,\varepsilon)}\right) & \text{if } d=0. \end{cases} \tag{19}$$

The values of d, m, and s depend on the model, and are listed in Table 1.

These 40 numbers in Table 1 replace several pages of tables, allowing the computation of critical values and p-values for the right tail of the distribution. The benefit of the approximation is that all replications enter the calculations, while 1% critical values would only be based on 100 for 10 000 replications. There are no further parameters in supF_{MAX} that need to be calibrated.

5 Simulations of tests for a single break at an unknown date

The size and power of the proposed test statistic is evaluated, and compared to three other procedures.

Experiments to evaluate the size are for DGP (8) under the null of no break and with a range of values for ρ while $\zeta = \theta = 0$ initially. Table 2 reports the empirical rejection frequencies at a 5% nominal size of the tests (14), (15), and (18). supF_{BPN} is listed in the last four columns of the table, which is the BP test without any corrections; supF_{BPQ} is in the middle, and uses the quadratic spectral method with standard prewhitening; the new test is in the first four columns (after the value of ρ). All tests use the same critical values from the distribution approximation for $T = 100$. As expected, supF_{BPN} only works for $\rho = 0$, although the size is somewhat on the low side at this sample size. supF_{BPN} works for small autoregressive values, but is already far from the nominal size for $\rho = 0.9$. supF_{MAX} has the correct size throughout, except that it is a bit high near 10% for models 2 and 3 in the I(1) case.

Model	0	1	2	3	0	1	2	3	0	1	2	3
ρ	$\sup F_{\text{MAX}}$				$\sup F_{\text{BPQ}}$				$\sup F_{\text{BPN}}$			
0.0	0.057	0.056	0.062	0.043	0.052	0.056	0.074	0.053	0.032	0.031	0.048	0.021
0.5	0.058	0.054	0.060	0.045	0.067	0.081	0.092	0.078	0.427	0.505	0.413	0.563
0.9	0.050	0.054	0.031	0.061	0.205	0.257	0.233	0.312	0.963	0.983	0.914	0.995
1.0	0.051	0.046	0.114	0.089	0.553	0.427	0.467	0.537	0.998	0.997	0.977	1.000

Table 2: Size of test for break at unknown date at 5% nominal size, $T = 100, \varepsilon = 0.1, 10\,000$ replications.

	Model	0	1	2	3
T	method	$\sup F_{\text{MAX}}$			
100	$p_{\text{app}}(100, 0.1)$	0.057	0.056	0.062	0.043
100	$p_{\text{app}}(1000, 0.1)$	0.071	0.070	0.083	0.067
1000	$p_{\text{app}}(1000, 0.1)$	0.051	0.054	0.058	0.040
1000	BP Table II	0.070	0.127	0.015	0.086

Table 3: Impact of choice of critical values on size of test for break at unknown date at 5% nominal size, $\rho = 0$,$\varepsilon = 0.1, 10\,000$ replications.

Table 3 takes a brief look at the impact of using different critical values. It confirms that the robustness adjustments do affect the small sample distribution, with the effect disappearing for $T = 1000$. The approximation (19) is tailored for $\sup F_{\text{MAX}}$, and the small sample adjustment helps controlling the size. The critical values of BP Table II (or Table I, as the first columns are the same) only apply to model 0, but still work fairly well in model 3. For model 0, the $0.1, 0.05, 0.01$ critical values from $p_{\text{app}}(\infty, 0.05)$ are $8.64, 10.13, 13.52$ respectively, and $9.05, 10.76, 14.73$ for $p_{\text{app}}(100, 0.05)$ This can be compared to $8.02, 9.63, 13.58$ in Table I of BP.

Table 4 adds three other existing procedures for comparison. PY is the test of Perron and Yabu (2009), HLT the test of Harvey, Leybourne, and Taylor (2010), and SV the J version of Sayginsoy and Vogelsang (2011) with the Daniell kernel. Critical values[3] for PY are the largest of the I(0) and I(1) entries in Perron and Yabu (2009, Table 2).

HLT has rather poor size control in Table 4, ranging from undersized with iid normal errors and oversized for I(1) errors. SV is also undersized for $\rho = 0$ but correctly sized for I(1) errors. PY is substantially oversized, except for model 2 close to the unit root, even though, as

[3] There is a difference between the published procedure and the GAUSS implementation at `blogs.bu.edu/perron/codes`: the former does the BIC lag length selection in the equation for the residuals, the latter in the equation for y_t. We follow the code, but do not copy a minor counting error in the computation of the test.

Model	1	2	3	1	2	3	2	3	2	3
ρ	supF_{MAX}			PY			HLT		SV	
	$\zeta = 0.0$ at 10% nominal size									
0.0	0.105	0.114	0.086	0.141	0.119	0.143	0.018	0.041	0.047	0.024
0.5	0.095	0.099	0.074	0.166	0.113	0.162	0.027	0.045	0.061	0.049
0.9	0.079	0.046	0.083	0.168	0.027	0.092	0.053	0.062	0.054	0.055
1.0	0.067	0.177	0.117	0.155	0.139	0.191	0.244	0.244	0.100	0.125
	$\zeta = 0.0$ at 5% nominal size									
0.0	0.056	0.062	0.043	0.087	0.065	0.096	0.009	0.020	0.016	0.009
0.5	0.054	0.060	0.045	0.109	0.065	0.117	0.016	0.025	0.022	0.017
0.9	0.054	0.031	0.061	0.098	0.020	0.069	0.036	0.040	0.026	0.021
1.0	0.046	0.114	0.089	0.085	0.086	0.136	0.175	0.175	0.056	0.060
	$\zeta = 0.0$ at 1% nominal size									
0.0	0.014	0.015	0.013	0.030	0.019	0.036	0.003	0.005	0.001	0.000
0.5	0.017	0.019	0.013	0.045	0.024	0.052	0.005	0.008	0.002	0.001
0.9	0.027	0.016	0.035	0.039	0.013	0.039	0.017	0.021	0.004	0.002
1.0	0.024	0.050	0.058	0.033	0.040	0.072	0.095	0.096	0.012	0.011
	$\zeta = 0.5$ at 5% nominal size									
0.0	0.011	0.000	0.004	0.077	0.051	0.103	0.002	0.007	0.002	0.000
0.5	0.012	0.000	0.004	0.074	0.043	0.105	0.008	0.011	0.035	0.011
0.9	0.000	0.002	0.001	0.057	0.012	0.027	0.022	0.024	0.015	0.024
1.0	0.003	0.105	0.017	0.054	0.123	0.152	0.239	0.240	0.034	0.076

Table 4: Size of test for break at unknown date, $T = 100, \varepsilon = 0.1, 10\,000$ replications.

recommended, we use the largest critical values here. Overall, it appears that supF_{MAX} is best behaved.

The bottom part of Table 4 introduces second order serial correlation. This makes supF_{MAX} undersized for models 2 and 3 (except I(1) model 2), so it would need adjustment to handle this better. For PY, SV and HLT, the rejection frequencies in the stationary models are mostly smaller than with $\zeta = 0$, but for the non-stationary cases the size moves in the opposite direction.

supF_{MAX} is the only test with correct size under the most basic case, i.e. with iid errors, which is combined with a size under I(1) errors that is reasonably close to nominal. The widely different size properties of the other procedures makes it more difficult to compare power. In Table 5 we use the fact that we have the complete tail of the distribution approximated. We select a nominal size of supF_{MAX} which gives an empirical rejection frequency under the null that is close to that observed for PY, SV, and HLT. This allows us to make pairwise comparisons, with supF_{MAX} effectively size corrected towards the other procedure. The adopted nominal size is listed in the rows with header α_c. The first part of the table has $\rho = 0$, and a trend break with coefficient 0.04. Model 3 also has a level shift of unity at the same point. We see that HLT has lower power, but otherwise the procedures are similar.

Model	2	3	2	3	2	3	2	3	2	3	2	3
	PY		supF_{MAX}		HLT		supF_{MAX}		SV		supF_{MAX}	
	I(0): $\rho = 0.0$, $\eta_1 = 0$ (model 2) or $\eta_1 = 1$ (model 3), $\eta_2 = 0.04$											
rejection	0.76	0.87	0.70	0.78	0.27	0.30	0.47	0.66	0.65	0.57	0.70	0.56
at α_c	*0.05*	*0.05*	*0.05*	*0.10*	*0.10*	*0.10*	*0.01*	*0.04*	*0.10*	*0.10*	*0.05*	*0.02*
	I(1): $\rho = 1$, $\eta_1 = 0$ (model 2) or $\eta_1 = 3$ (model 3), $\eta_2 = 0.5$											
rejection	0.38	0.63	0.35	0.35	0.53	0.50	0.53	0.42	0.18	0.28	0.35	0.31
at α_c	*0.01*	*0.01*	*0.01*	*0.02*	*0.01*	*0.01*	*0.04*	*0.05*	*0.05*	*0.05*	*0.01*	*0.01*

Table 5: Power of test for break in trend at unknown date at matching nominal size α_c, $T = 100, \varepsilon = 0.1, 10\,000$ replications, $\zeta = 0$. Break in DGP at $t = 50$.

The next part of Table 5 is for $\rho = 1$. Then it is much more difficult to detect a trend, and the DGP parameters are adjusted accordingly. Two entries stand out: SV has low power for model 2, and PY has high power for model 3.

6 Multiple break tests

Testing for multiple breaks commences with testing for no break, and, if that is rejected, proceeds to the hypothesis of two breaks given one. We argue that the asymptotic distributions that are tabulated in the literature do not correspond to the proposed procedure. We also simulate the properties of test procedures for up to three breaks in the DGP.

We limit ourselves to the expanding sequence of tests for m breaks given $m-1$ breaks, denoted $F(m|m-1)$. The adopted approach is proposed by Bai and Perron (1998), which can be described without being specific about which version is used for $F(1)$, the test for a single break.

In each case, the first objective is to set M as the largest number of breaks to consider, and estimate the break sets that minimize RSS: $\widehat{\mathcal{B}}(1),...,\widehat{\mathcal{B}}(M)$. Section 2 discussed procedures to obtain these sets, and we use the proposed algorithm in the results below. Note that, when moving from one to two breaks (say), the set is not conditional on the single break, but re-estimated for two breaks. This is particularly useful if the first set is too small and wrongly estimated, but the test nonetheless rejects the null hypothesis.

For the sequence of tests, BP propose a sample splitting algorithm: for m breaks, cut the sample at the optimal break set for $m-1$ breaks. Then compute test $F(1)$ in each subsample; if there is no space to insert a further break, that segment is ignored. The maximum of these m tests is the statistic of interest, $F(m|m-1)$. The convenience of this approach is lost if the expression for $F(1)$ is made to depend on m.

The asymptotic distribution for this sequential procedure is shown to be the m-th power of the test for a single break. This asymptotic result assumes that ε is kept fixed. In practice, however, the minimum sample size (i.e. the gap between breaks), is kept fixed as the sample shrinks, so that ε grows accordingly. This is how the dynamic programming algorithm BP03 works, and also adopted in the proposed algorithm (although this could be changed). Further discussion of this issue is given in Appendix B, together with normal approximations to the distributions when $F(1) = \text{sup}F_{\text{MAX}}$.

The simulation experiments allow for up to three breaks, equally spaced in the sample, using DGP (9) with $\zeta = \theta = 0$ and a range of values for ρ. With M^* denoting the number of breaks in the DGP, we use:

$$\begin{array}{lll}
\text{Model 0:} & \boldsymbol{\nu}_1 = 3\boldsymbol{\iota}_{M^*}, & \boldsymbol{\nu}_2 = \mathbf{0}, \\
\text{Model 1:} & \boldsymbol{\nu}_1 = 3\boldsymbol{\iota}_{M^*}, & \boldsymbol{\nu}_2 = \mathbf{0}, \\
\text{Model 2:} & \boldsymbol{\nu}_1 = 0, & \boldsymbol{\nu}_2 = 0.5\boldsymbol{\iota}_{M^*}, \\
\text{Model 3:} & \boldsymbol{\nu}_1 = 3\boldsymbol{\iota}_{M^*}, & \boldsymbol{\nu}_2 = 0.5\boldsymbol{\iota}_{M^*},
\end{array}$$

where $\boldsymbol{\iota}_{M^*}$ is a vector of ones of length M^*. The optimal break set is found with the new algorithm, and we apply the sequential procedure for testing.

Two additional procedures are added to the simulations. The first is the Kejriwal and Perron (2010) extension of PY to sequential testing, denoted PY-KP. We use the PY procedure as described in Perron and Yabu (2009), together with the largest of the I(0) and I(1) critical values in Kejriwal and Perron (2010, Table 1). The second is the Sobreira and Nunes (2016) extension of Harvey, Leybourne, and Taylor (2010), denoted HLT-SN. We deviate from SN in that we do not adopt their adjustments to the test. Instead we use the original HLT procedure, so keep the

		supF_{MAX}				PY-KP				HLT-SN			
ρ	T	1|0	2|1	3|2	4|3	1|0	2|1	3|2	4|3	1|0	2|1	3|2	4|3
		$M^* = 1$: 1 break in DGP											
0.0	60	*1.00*	0.06	0.01	0.00	*1.00*	0.10	0.03	0.01	*0.97*	0.00	0.00	0.00
0.5	60	*0.91*	0.07	0.02	0.01	*0.80*	0.11	0.05	0.02	*0.90*	0.00	0.00	0.00
0.9	60	*0.37*	0.09	0.04	0.02	*0.38*	0.13	0.08	0.04	*0.58*	0.07	0.02	0.01
1.0	60	*0.43*	0.11	0.05	0.02	*0.46*	0.15	0.08	0.04	*0.57*	0.11	0.03	0.01
0.0	120	*1.00*	0.06	0.01	0.00	*1.00*	0.06	0.01	0.00	*1.00*	0.00	0.00	0.00
0.5	120	*1.00*	0.06	0.01	0.00	*1.00*	0.06	0.02	0.00	*1.00*	0.01	0.00	0.00
0.9	120	*0.75*	0.06	0.03	0.01	*0.77*	0.05	0.03	0.01	*0.93*	0.05	0.02	0.01
1.0	120	*0.68*	0.09	0.03	0.02	*0.70*	0.06	0.03	0.02	*0.75*	0.10	0.03	0.01
		$M^* = 2$: 2 breaks in DGP											
0.0	60	*0.97*	*0.98*	0.04	0.01	*0.99*	*0.97*	0.07	0.02	*1.00*	*0.14*	0.00	0.00
0.5	60	*0.82*	*0.39*	0.05	0.01	*0.88*	*0.44*	0.09	0.02	*0.99*	*0.10*	0.00	0.00
0.9	60	*0.80*	*0.11*	0.06	0.02	*0.91*	*0.15*	0.10	0.05	*0.95*	*0.09*	0.03	0.01
1.0	60	*0.78*	*0.11*	0.06	0.02	*0.85*	*0.16*	0.11	0.04	*0.89*	*0.12*	0.04	0.01
0.0	120	*0.98*	*1.00*	0.05	0.01	*1.00*	*1.00*	0.04	0.00	*1.00*	*0.98*	0.00	0.00
0.5	120	*1.00*	*0.97*	0.05	0.01	*1.00*	*0.91*	0.04	0.01	*1.00*	*0.84*	0.00	0.00
0.9	120	*1.00*	*0.19*	0.05	0.02	*1.00*	*0.14*	0.05	0.02	*1.00*	*0.27*	0.03	0.01
1.0	120	*0.98*	*0.17*	0.05	0.02	*0.99*	*0.12*	0.05	0.02	*0.99*	*0.23*	0.05	0.01
		$M^* = 3$: 3 breaks in DGP											
0.0	60	*0.99*	*1.00*	*0.48*	0.01	*1.00*	*0.99*	*0.59*	0.01	*1.00*	*0.39*	*0.01*	0.00
0.5	60	*0.99*	*0.57*	*0.11*	0.01	*1.00*	*0.63*	*0.18*	0.02	*1.00*	*0.33*	*0.01*	0.00
0.9	60	*0.98*	*0.21*	*0.04*	0.02	*1.00*	*0.29*	*0.08*	0.04	*1.00*	*0.28*	*0.02*	0.01
1.0	60	*0.95*	*0.21*	*0.04*	0.02	*0.98*	*0.28*	*0.08*	0.04	*0.99*	*0.29*	*0.02*	0.01
0.0	120	*1.00*	*1.00*	*1.00*	0.00	*1.00*	*1.00*	*0.99*	0.00	*1.00*	*1.00*	*0.39*	0.00
0.5	120	*1.00*	*0.99*	*0.64*	0.00	*1.00*	*0.95*	*0.51*	0.00	*1.00*	*0.96*	*0.28*	0.00
0.9	120	*1.00*	*0.44*	*0.06*	0.01	*1.00*	*0.35*	*0.07*	0.01	*1.00*	*0.62*	*0.06*	0.00
1.0	120	*1.00*	*0.42*	*0.05*	0.02	*1.00*	*0.36*	*0.05*	0.01	*1.00*	*0.54*	*0.05*	0.01

Table 6: Model 2 rejection frequency of tests $F(m|m-1)$ for up to three breaks in the DGP; 5% nominal size, $T = 60, 120, \varepsilon = 0.15, 10\,000$ replications.

weight adjustment (denoted b_ξ by SN) fixed at the HLT values, and also do not make the weights themselves a function of m. We do use the critical values of Sobreira and Nunes (2016, Table 2 and 4), noting that they are based on only 5000 replications.

The sample sizes for testing are 60 and 120. This is in contrast to the larger samples used in the simulations of KP (120,240,360) and SN (150). The results for model 2 are in Table 6 and for model 3 in Table 7. The tables use $\varepsilon = 0.15$, because that is the only value considered by SN. The tables report the rejection frequency of testing up to four breaks when the DGP has from one to three breaks at $\alpha_c = 5\%$. Italic columns correspond to a break in the DGP.

For model 2 we see that PY-KP is relatively oversized. As before, the size of HLT-SN when testing for one break is sensitive to the value of ρ, but seems to have some power advantage for

		supF_{MAX}				PY-KP				HLT-SN			
ρ	T	1|0	2|1	3|2	4|3	1|0	2|1	3|2	4|3	1|0	2|1	3|2	4|3
		$M^* = 1$: 1 break in DGP											
0.0	60	*1.00*	0.05	0.01	0.00	*1.00*	0.23	0.07	0.03	*0.93*	0.00	0.00	0.00
0.5	60	*0.91*	0.08	0.02	0.01	*0.91*	0.27	0.13	0.05	*0.84*	0.00	0.00	0.00
0.9	60	*0.41*	0.12	0.04	0.01	*0.70*	0.31	0.19	0.08	*0.53*	0.06	0.01	0.00
1.0	60	*0.41*	0.14	0.05	0.02	*0.72*	0.34	0.21	0.08	*0.52*	0.10	0.02	0.00
0.0	120	*1.00*	0.05	0.01	0.00	*1.00*	0.11	0.02	0.00	*1.00*	0.00	0.00	0.00
0.5	120	*1.00*	0.07	0.01	0.00	*1.00*	0.14	0.04	0.01	*1.00*	0.01	0.00	0.00
0.9	120	*0.55*	0.11	0.05	0.02	*0.86*	0.13	0.08	0.03	*0.92*	0.05	0.01	0.00
1.0	120	*0.47*	0.13	0.06	0.02	*0.83*	0.17	0.10	0.03	*0.74*	0.12	0.03	0.00
		$M^* = 2$: 2 breaks in DGP											
0.0	60	*0.66*	*0.95*	0.04	0.01	*0.87*	*0.98*	0.20	0.06	*0.97*	*0.09*	0.00	0.00
0.5	60	*0.35*	*0.42*	0.06	0.01	*0.86*	*0.66*	0.22	0.08	*0.95*	*0.06*	0.00	0.00
0.9	60	*0.40*	*0.22*	0.08	0.02	*0.94*	*0.52*	0.26	0.12	*0.88*	*0.07*	0.03	0.00
1.0	60	*0.45*	*0.22*	0.08	0.02	*0.93*	*0.54*	0.26	0.12	*0.83*	*0.09*	0.03	0.00
0.0	120	*0.76*	*1.00*	0.04	0.00	*1.00*	*1.00*	0.08	0.01	*1.00*	*0.96*	0.00	0.00
0.5	120	*0.46*	*0.97*	0.05	0.01	*1.00*	*0.96*	0.10	0.03	*1.00*	*0.81*	0.00	0.00
0.9	120	*0.59*	*0.38*	0.08	0.03	*1.00*	*0.52*	0.11	0.05	*1.00*	*0.22*	0.02	0.00
1.0	120	*0.71*	*0.29*	0.08	0.04	*0.99*	*0.49*	0.13	0.05	*0.98*	*0.20*	0.04	0.01
		$M^* = 3$: 3 breaks in DGP											
0.0	60	*0.73*	*0.95*	*0.50*	0.00	*0.99*	*0.99*	*0.76*	0.01	*1.00*	*0.10*	*0.02*	0.00
0.5	60	*0.43*	*0.54*	*0.13*	0.00	*0.99*	*0.84*	*0.43*	0.03	*1.00*	*0.12*	*0.02*	0.00
0.9	60	*0.43*	*0.33*	*0.07*	0.01	*1.00*	*0.70*	*0.38*	0.04	*0.99*	*0.17*	*0.01*	0.00
1.0	60	*0.49*	*0.32*	*0.08*	0.01	*0.99*	*0.70*	*0.40*	0.05	*0.97*	*0.19*	*0.01*	0.00
0.0	120	*0.97*	*1.00*	*1.00*	0.00	*1.00*	*1.00*	*0.98*	0.00	*1.00*	*0.90*	*0.45*	0.00
0.5	120	*1.00*	*0.99*	*0.63*	0.00	*1.00*	*0.98*	*0.62*	0.01	*1.00*	*0.79*	*0.26*	0.00
0.9	120	*0.96*	*0.57*	*0.20*	0.02	*1.00*	*0.76*	*0.31*	0.02	*1.00*	*0.50*	*0.06*	0.00
1.0	120	*0.93*	*0.49*	*0.17*	0.02	*1.00*	*0.73*	*0.31*	0.03	*1.00*	*0.47*	*0.05*	0.00

Table 7: Model 3 rejection frequency of tests $F(m|m-1)$ for up to three breaks in the DGP; 5% nominal size, $T = 60, 120, \varepsilon = 0.15, 10\,000$ replications.

larger ρ. When testing for two breaks given one, when there are indeed two breaks we see that HLT-SN has no power at $T = 60$, recovers at $T = 120$. Arguable $T = 60$ is more relevant for annual data. The same happens for three breaks when there are three, but now it persists for $T = 120$.

Finally, we look at the rejection frequencies for different values of ε: Table 8 shows $\varepsilon = 0.1$ and $\varepsilon = 0.05$ when there are two breaks in the DGP; Table 7 has $\varepsilon = 0.15$. The rejection frequency is lower for smaller ε. This is particularly pronounced for PY-KP, where too many breaks are accepted.

Our preference from these results tends towards supF_{MAX}.

		supF_{MAX}						PY-KP					
		model 2			model 3			model 2			model 3		
ρ	ε	$2\|1$	$3\|2$	$4\|3$	$2\|1$	$3\|2$	$4\|3$	$2\|1$	$3\|2$	$4\|3$	$2\|1$	$3\|2$	$4\|3$
		$M^* = 2$: 2 breaks in DGP											
0.0	0.05	*1.00*	0.06	0.04	*0.70*	0.06	0.03	*1.00*	0.11	0.08	*1.00*	0.27	0.21
0.5	0.05	*0.97*	0.07	0.06	*0.34*	0.08	0.07	*0.94*	0.14	0.14	*0.98*	0.33	0.34
0.9	0.05	*0.17*	0.11	0.12	*0.52*	0.17	0.17	*0.18*	0.19	0.22	*0.61*	0.42	0.51
1.0	0.05	*0.17*	0.14	0.12	*0.66*	0.21	0.21	*0.18*	0.21	0.23	*0.59*	0.49	0.55
0.0	0.10	*1.00*	0.05	0.01	*0.74*	0.06	0.01	*1.00*	0.08	0.02	*1.00*	0.20	0.06
0.5	0.10	*0.98*	0.07	0.03	*0.40*	0.08	0.02	*0.94*	0.10	0.04	*0.98*	0.26	0.13
0.9	0.10	*0.19*	0.09	0.05	*0.56*	0.15	0.08	*0.16*	0.13	0.08	*0.59*	0.29	0.22
1.0	0.10	*0.18*	0.10	0.05	*0.69*	0.17	0.10	*0.16*	0.13	0.08	*0.56*	0.33	0.24

Table 8: Model 2 and 3 rejection frequency of tests $F(m|m-1)$ for up to three breaks in the DGP; 5% nominal size, $T = 120, 10\,000$ replications.

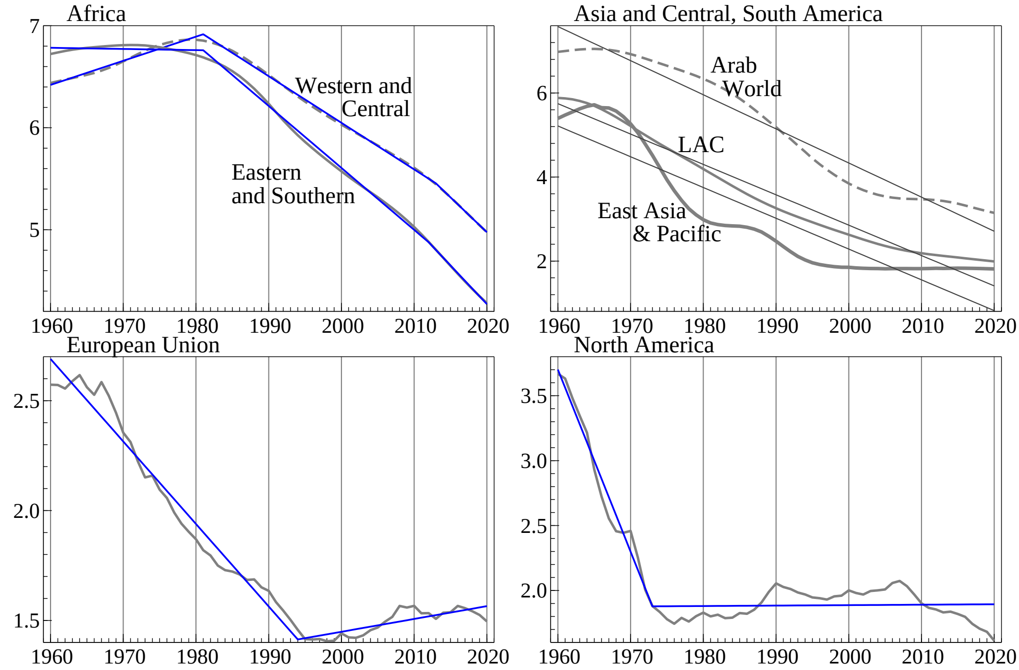

Figure 7: Trend estimates of fertility for 6 regions of the world. LAC is Latin America & Caribbean. Model 2 supF_{MAX} tests with $\varepsilon = 0.1, \alpha_c = 0.01$.

7 Testing for breaks in fertility data

We test for broken trends in fertility using data from the World Bank, considering annual data for the period 1960 to 2020.

The fertility data is taken from the World Bank database (indicator SP.DYN.TFRT.IN), which provides annual data from 1960 to 2020 for most countries, as well as regions of the globe, and some subdivisions by income groups. However, in a subset of domains the data consists of lower frequency observations with interpolation in between. The fertility rate is defined as the 'the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year'. Preliminary visual inspection suggests that there has been a noticeable change in the underlying trends, at least in several countries.

The analysis is restricted to 192 individual countries with data, 12 regions of the world, and the entire world. We first note that with supF_{BPN} all but 10 countries appear to have at least two trend breaks when testing at $\alpha_c = 1\%$. This reduces to 46 with supF_{BPQ}, and 27 with supF_{MAX}. We adopted $\varepsilon = 0.1$ to have at least 6 observations in a partial trend, and allow up to three breaks. The remaining results all use supF_{MAX}. We use model 2 for testing, because a simultaneous level and trend break seems implausible in this case.

Figure 7 shows the estimated trends for 7 of the 13 regions. Four have breaks: the null hypothesis of no break is rejected. In the case of Africa there are two breaks in each, although

	Germany		United Kingdom	
Start	supF_{MAX}	p-value	supF_{MAX}	p-value
1960	6.3150	0.083	4.8432	0.158
1961	7.7094	0.046	8.0737	0.039
1962	10.224	0.016	12.639	0.006
1963	14.142	0.004	19.988	0.000
1964	18.320	0.001	23.809	0.000
1965	28.428	0.000	23.535	0.000
1966	40.155	0.000	22.383	0.000

Table 9: Model 2 tests for one break in fertility using different starting dates for the estimation sample.

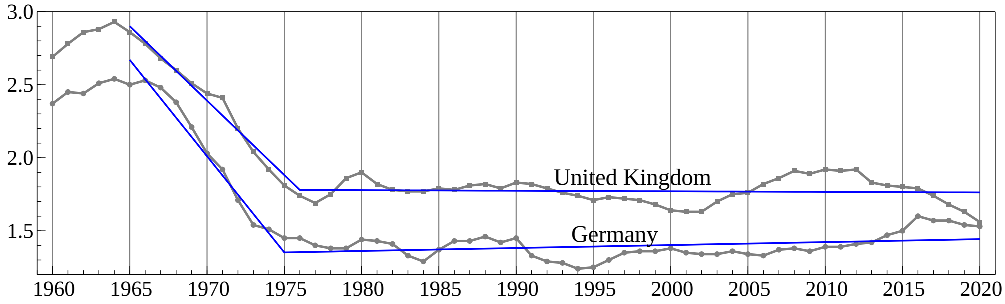

Figure 8: Break found in fertility for the UK and Germany when discarding the initial 5 observations.

the data looks heavily interpolated. They appear I(2) and the trends approximate the curves; Western and central Africa shows a trend reversal around 1981.

The downward trend for European Union did not start until about 1967. Several European countries show a few years of upward trend, before a rapid decline in fertility during the second half of the 1960s and the 1970s. This data feature plays havoc with the test: the initial segment is too short to be considered as a separate trend with $\varepsilon = 0.1$. When reducing ε to 0.05, the new RSS minimization algorithm puts the dates for one and two breaks in the right place, but the test for one break is insignificant. Table 9 moves the start of the sample one observation forward at a time. It shows how the test suddenly becomes highly significant. Seven European countries exhibit this effect: Austria, France, Germany, Italy, Luxembourg, Norway, and the UK. This is a situation where some judgment is required, see Figure 8, which depicts the fertility in the UK and Germany, starting from 1965.

If we look in more detail at the eight largest EU countries, we find that half have a break, see Figure 9. Visually, the downward trends in fertility started in the early to mid 1960s. That period saw many changes in society, as well as major advances in contraceptive technology, giving more control to women of reproductive age. The first oral contraceptive was approved in the United Status in 1960 (only for married women until 1972), and 1962 in the United Kingdom; by 1968

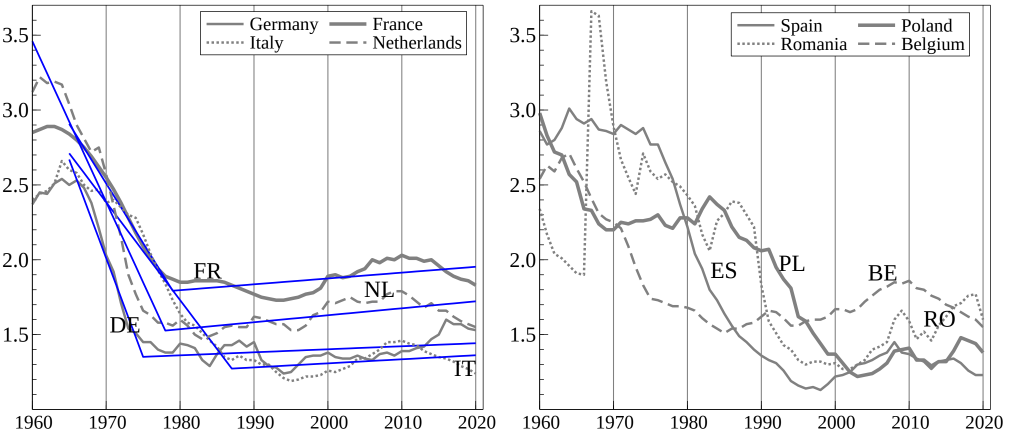

Figure 9: Fertility of 8 largest EU countries, with break on the left, no break found on the right.

each had approved eight first-generation oral contraceptives, see Gelijns and Pannenborg (1993). In addition, the mid 1960s and 1970s saw rapid development of intrauterine devices, while surgical procedures also improved. Mauldin and Segal (1988) find that the fertility rate declines with contraceptive prevalence, to the extent that a 14 percent increase in married women of child-bearing age practicing contraception reduces fertility by one child. This effect is somewhat lower for the crude birth rate, which they consider an inferior measure because it depends on the age structure of the population. However, we also see a well-defined end to the downward trends in fertility, which could mean that the previously unmet demand for contraception was largely satisfied after a period of 10 to twenty years.

Trend breaks are less prominent in the crude birth rates (World Bank id SP.DYN.CBRT.IN), defined as the number of live births occurring during the year per 1000 population at midyear. While the secular patterns are very similar, the break is only significant for the European Union, Figure 10. For the eight largest EU countries the first change is the Netherlands, where the break now has a p-value of 1.6%. As happened before, starting five observations later reduces the p-value to 0.02%. The other change is that for France two breaks given one is not rejected, as can be seen from Figure 11. For Italy the second break is marginal with a p-value at 2.2%.

In general, the statistical tests suggest a significant downwards trend from the mid to early 1960's through the 1970s in many western countries, followed by a break and pronounced levelling off.

8 Conclusions

We studied test procedures for multiple trend breaks. The specification where the trend breaks, in the presence of a fixed level was deemed more empirically relevant than the model in which there is a level shift and local trend changing at the same time. An algorithm to minimize the residual sum of squares was developed, which, in particular, is valid for the former specification,

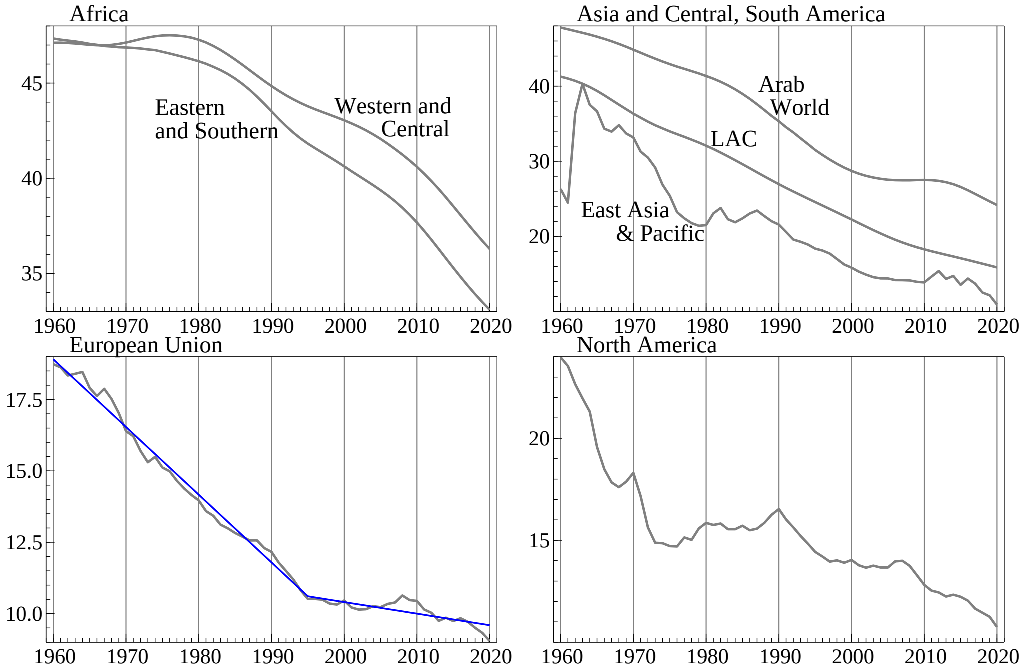

Figure 10: Trend estimates of crude birth rates for 6 regions of the world. LAC is Latin America & Caribbean. Model 2 supF_{MAX} tests with $\varepsilon = 0.1, \alpha_c = 0.01$.

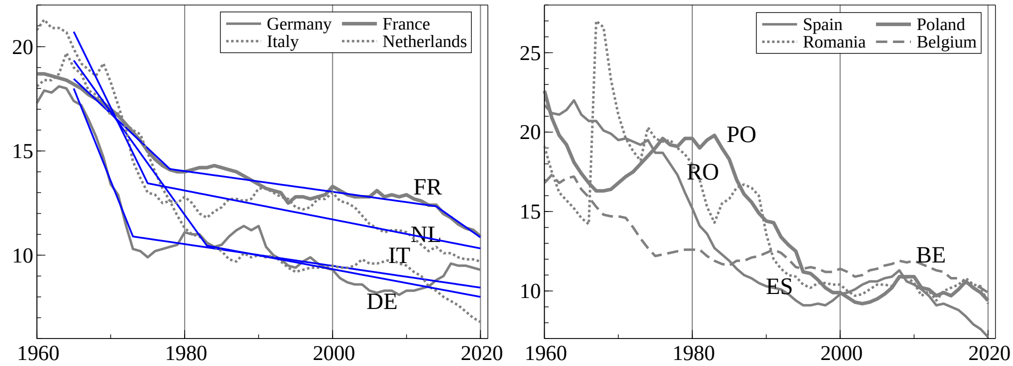

Figure 11: Crude birth rates of 8 largest EU countries, with break on the left, no break found on the right.

but can be used in the other cases as well. We introduced a test that can be seen as a simplified version of an existing test, and provided an approximation to the distribution for small and large sample sizes. The test was shown to have decent small sample properties, even when the errors are I(1).

The focus of this paper is mainly on practical aspects to help implementation and empirical applications. It is asserted that the approach fits in the asymptotic frameworks of the existing literature, but leave formal analysis to subsequent work.

We studied fertility for all countries of the world using the new test, with significant changes in trend found in four of the largest countries of the European Union, as well as Canada, the United Kingdom, and the United States. For high fertility countries, the data is usually interpolated, and too smooth. That pattern often resulted in an upward trend, followed by one or more downward ones. The method employed imposes a minimum length of the trend, often set to ten or fifteen percent of the sample. For fertility that corresponds to six or nine years, so it will be a while before we can study the impact of the Covid pandemic in this way.

A An algorithm to find break dates

The first requirement is to get an efficient method to estimate model (2) by OLS at all possible break points s. Start by writing it in matrix form as

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{X}(s)\boldsymbol{\psi}(s) + \boldsymbol{u}(s). \tag{20}$$

Where $\boldsymbol{D}$ has q columns holding $\boldsymbol{d}_t$ as well as all additional non-breaking regressors. $\boldsymbol{X}(s)$ is $T \times k$. Take the thin QR decomposition (Golub and Van Loan, 2013, Ch.5):

$$\boldsymbol{D}\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{P}$ is the matrix that rotates columns of $\boldsymbol{D}$, $\boldsymbol{Q}$ is of dimension $T \times q$ and orthogonal such that $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_q$, and $\boldsymbol{R}$ is $q \times q$ with zeros below the diagonal. Then the transformed system is:

$$\boldsymbol{Q}'\boldsymbol{y} = \boldsymbol{R}\boldsymbol{P}'\boldsymbol{\beta} + \boldsymbol{Q}'\boldsymbol{X}(s)\boldsymbol{\psi}(s) + \widetilde{\boldsymbol{u}}.$$

This shows that the (reordered) $\boldsymbol{\beta}$ coefficients are solved from the first q observations of the transformed system, with residuals zero. (For reduced rank regressors q is reduced.) The $\boldsymbol{\psi}(s)$ coefficients are estimated from the remaining observations in a regression on $\boldsymbol{Q}'\boldsymbol{X}(s)$. Because we are only interested in the RSS, we can compute that directly from this regression, without the need to undo the pivoting or transformation. So with this approach we can relatively efficiently solve:

$$\widehat{s} = \underset{s\in[T_0,T_1]}{\arg\min} \left\{ \text{RSS}(s) |\, \boldsymbol{y} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{X}(s)\boldsymbol{\psi}(s) + \boldsymbol{u}(s) \right\}. \tag{21}$$

So, while we require a regression at each break point, this only has k regressors, regardless of q ($k = 1$ except for model 3 where it is 2). This allows us to formulate the core of the new algorithm:

Core[$\mathcal{B}_j(m), G$]:

Given a set of m breaks $\mathcal{B}_j(m)$ do for each break $i = 1, ..., m$:

1. Construct a set without break i: $\mathcal{B}_j(m\backslash i)$.
2. Find index $\widehat{s}_i$ from solving (21) conditional on breaks $\mathcal{B}(m\backslash i)$, only allowing break dates that are more than G observations away from the dates in $\mathcal{B}_j(m\backslash i)$.

```
3. Insert $\widehat{s}_i$ at position $i$ to construct the new break set $\mathcal{B}_{j+1}(m)$.
```

Unlike the BP03 algorithm, it is not straightforward to satisfy spacing between breaks. Because the objective function is not smooth, we gradually grow G during iteration until it is sufficiently large.

New algorithm:

```
1. Solve (21) to find the first break, and so $\widehat{\mathcal{B}}(1)$.          (initialization)
2. For $m = 2, ..., M$                                  (loop over expanding break sets)
   (a) Set $\mathcal{B}_0(m) = \{T, \widehat{\mathcal{B}}(m-1)\}$, $G = 1$
   (b) For $j = 1, ..., J$ :                              (iteration over this break set)
         $\mathcal{B}_j(m) = \text{Core}[\mathcal{B}_j(m), G]$                    (update break set)
       If $G < \lfloor \varepsilon T \rfloor$ set $G = \min\{G + 2, \lfloor \varepsilon T \rfloor\}$ and continue with next $j$.
       If $|\text{RSS}[\mathcal{B}_{j-1}(m) - \text{RSS}[\mathcal{B}_j(m)]| \le \epsilon_1 \text{RSS}[\mathcal{B}_j(m)]$ continue with next $m$, unless there is no space left between breaks.
       Otherwise continue with next $j$.
```

Convergence requires that the desired gap between breaks is reached, and no further progress is made. We set $\epsilon_1 = 10^{-12}$ and maximum number of iterations $J = 20$.

B The role of ε in the sequential test and approximations to the distribution

The ε value controls the spacing between breaks, and how this is handled matters in the sequential test with sample splitting. It is perhaps most easily explained with an example using $\varepsilon = 0.1$ and observations $t = 1, ..., T = 100$. Then $\varepsilon T = 10$ and RSS is minimized with the break date candidates in $\varepsilon T, ..., T - \varepsilon T + 1$, so $10, ..., 91$ here. In other words, for RLS the smallest segment has 10 observations.

For the next step, assume we established a break at $\widehat{\mathcal{B}}(1) = \{40\}$, and move to the test for two breaks. We evaluate the test in the first sample, $t = 1, ..., 40$, and in the second, $t = 41, ..., 100$. Now the choice of ε in these two smaller samples matters. The RSS minimization algorithm enforces a separation of $\varepsilon T = 10$, so if the break at forty survives, the two-break set has the form $\widehat{\mathcal{B}}(2) = \{40, A\}$ where A cannot fall on observations $1, ..., 9, 32, ..., 49, 92, ..., 100$. So, when we compute the test for one break in each subsample, the value of ε needs to be adjusted to respect this.

This matters in practice and in theory. If there is a second break before 40 closer to the end than the gap allows, and the test uses $\varepsilon = 0.1$ in the subsample (i.e a smallest segment of 4), then this date will determine the test outcome. However, it is not a date that can be reached. In that case, the sequential test will keep rejecting more and more breaks until it runs out of space.

The solution is to adjust the effective ε used in each subsample to respect the minimum segment size of εT. This is automatically done when the BP03 algorithm or the proposed algorithm is used to provide the date for the test. However, while the Bai & Perron GAUSS code works this way, the literature tends to differ. E.g. Kejriwal and Perron (2010, §3.1) keeps ε fixed. This does affect the asymptotic distribution, suggesting that the published tables do not apply to the sequential procedure when using the suggested break detection algorithms.

The same procedure as in §4 is used to simulate the distribution of the sequential test $F(m|m-1)$. The test $F(2|1)$ is simulated under the null by splitting the sample in half, and taking the maximum of the 1-break test in each sample using cut-off 2ϵ. Then for $F(3|1)$ the sample is split in three, and the maximum taken for each subsample with cut-off 3ϵ, etc. A response surface is fitted to the means and standard deviations from the transformed test at sample sizes $60, 100, 125, 150, 200, 250, 500, 1000$ for ε in $\{0.01, 0.03, 0.05, 0.07, 0.10\}$ using $10\,000$ replications, provided $\lfloor \varepsilon T \rfloor \geq 5$. $F(5|4)$ excludes $\varepsilon = 0.1$. The results are in Table 10.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*, 817–858.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*, 821–856.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*, 47–78.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics 18*, 1–22. doi:10.1002/jae.659.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2020). Trend-indicator saturation. mimeo, Department of Economics, Oxford University.

Doornik, J. A. (1998). Approximations to the asymptotic distribution of cointegration tests. *Journal of Economic Surveys 12*, 573–593. Reprinted in M. McAleer and L. Oxley (1999). *Practical Issues in Cointegration Analysis*. Oxford: Blackwell Publishers.

Doornik, J. A. (2021). *Object-Oriented Matrix Programming using Ox* (9th ed.). London: Timberlake Consultants Press.

Engler, E. and B. Nielsen (2009). The empirical process of autoregressive residuals. *Econometrics Journal 12*, 367–381.

Garcia, R. and P. Perron (1996). An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics 78*, 111–25.

Gelijns, A. C. and C. O. Pannenborg (1993). The development of contraceptive technology. *International Journal of Technology Assessment in Health Care 9*, 210–232.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations* (4th ed.). Baltimore: The Johns Hopkins University Press.

Harvey, D. I., S. J. Leybourne, and A. M. R. Taylor (2010). Robust methods for detecting multiple level breaks in autocorrelated time series. *Journal of Econometrics 157*, 342–358.

Kejriwal, M. and P. Perron (2010). A sequential procedure to determine the number of breaks in trend with an integrated or stationary noise component. *Journal of Time Series Analysis 31*, 305–328. doi.org/10.1111/j.1467-9892.2010.00666.x.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1993). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics 54*, 159–178.

Mauldin, W. P. and S. J. Segal (1988). Prevalence of contraceptive use: Trends and issues. *Studies in Family Planning 19*, 335–353.

Nielsen, B. (1997). Bartlett correction of the unit root test in autoregressive models. *Biometrika 84*, 500–504.

Perron, P. and T. Yabu (2009). Testing for shifts in trend with an integrated or stationary noise component. *Journal of Business and Economic Statistics 27*, 369–396. doi.org/10.1198/jbes.2009.07268.

Roy, A. and W. A. Fuller (2001). Estimation for autoregressive processes with a root near one. *Journal of Business and Economic Statistics 19*, 482–493.

Sayginsoy, Ö. and T. J. Vogelsang (2011). Testing for a shift in trend at an unknown date: a fixed-B analysis of heteroskedasticity autocorrelation robust OLS-based tests. *Econometric Theory 27*, 992–1025.

Sobreira, N. and L. C. Nunes (2016). Tests for multiple breaks in the trend with stationary or integrated shocks. *Oxford Bulletin of Economics and Statistics 78*, 394–411. doi: 10.1111/obes.12116.

Yang, J. (2017). Consistency of trend break point estimator with underspecified break number. *Econometrics 5*. doi: 10.3390/econometrics5010004.

			Coefficients for m and s equations					
Model	d	m, s	$100/T$	$100^2/T^2$	ε	$\varepsilon^{1/2}$	$\varepsilon 100/T$	1
2 breaks given 1								
0	0	$m(T,\varepsilon) =$	0.00954	0.00398	0	−0.120	−0.0814	1.22
0	0	$s(T,\varepsilon) =$	0.0102	0	0	0.0363	0	0.0418
1	0	$m(T,\varepsilon) =$	0.0770	0.0489	0	−0.593	−0.681	2.06
1	0	$s(T,\varepsilon) =$	0.0931	0	0	0.189	0	0.320
2	0.30	$m(T,\varepsilon) =$	0.0582	0	0	−0.499	−0.213	1.36
2	0.30	$s(T,\varepsilon) =$	0.0138	0.00973	0	0.137	0	0.126
3	0	$m(T,\varepsilon) =$	0.0936	0.0415	−2.26	0	0	2.30
3	0	$s(T,\varepsilon) =$	0.0750	0.0140	0	0.271	0.119	0.256
4 breaks given 3								
0	0	$m(T,\varepsilon) =$	0.187	0.0875	0	−1.15	−1.85	2.08
0	0	$s(T,\varepsilon) =$	0.0976	0.0214	0.419	0.191	0	0.324
1	0	$m(T,\varepsilon) =$	0.210	0.0684	−1.83	0	−1.34	2.10
1	0	$s(T,\varepsilon) =$	0.107	0.0211	0	0.242	0	0.284
2	0.20	$m(T,\varepsilon) =$	0.103	0	−0.479	−0.406	−0.361	1.40
2	0.20	$s(T,\varepsilon) =$	0.0280	0.0164	0.150	0.0930	0	0.118
3	0	$m(T,\varepsilon) =$	0.296	0.0638	−2.91	0	−0.955	2.37
3	0	$s(T,\varepsilon) =$	0.133	0.0243	0.838	0	0	0.251
3 breaks given 2								
0	0	$m(T,\varepsilon) =$	0.229	0.159	−3.48	0	−2.62	2.06
0	0	$s(T,\varepsilon) =$	0.141	0.0219	0.982	0	0	0.317
1	0	$m(T,\varepsilon) =$	0.327	0.110	−6.91	1.72	−2.40	2.03
1	0	$s(T,\varepsilon) =$	0.168	0.0146	1.57	−0.371	0	0.317
2	0.15	$m(T,\varepsilon) =$	0.125	0.0160	−1.72	0	−0.563	1.39
2	0.15	$s(T,\varepsilon) =$	0.0562	0.0222	0.464	0	−0.165	0.118
3	0	$m(T,\varepsilon) =$	0.490	0.102	−7.32	1.42	−2.28	2.30
3	0	$s(T,\varepsilon) =$	0.165	0	0.834	0	0.757	0.243
5 breaks given 4								
0	0	$m(T,\varepsilon) =$	0.0870	0.0829	−1.16	0	−0.957	1.53
0	0	$s(T,\varepsilon) =$	0.0643	0.0280	0.234	0	−0.237	0.0929
1	0	$m(T,\varepsilon) =$	0.0633	0.0235	−0.356	0	−0.531	1.25
1	0	$s(T,\varepsilon) =$	0.0235	0.00823	0.0652	0	0	0.0349
2	0.10	$m(T,\varepsilon) =$	0.193	0	−2.05	0	−0.708	1.40
2	0.10	$s(T,\varepsilon) =$	0.0494	0.0462	0	0.176	0	0.102
3	0	$m(T,\varepsilon) =$	0.711	0.139	−4.39	0	−3.47	2.46
3	0	$s(T,\varepsilon) =$	0.195	0.0366	1.02	0	0.623	0.232

Table 10: Parameters m, a for the approximation to the mean and standard deviation of the transformed statistic for testing $F(m|m-1), m = 2, ..., 5$ using supF_{MAX}.