Current Trends in the Analysis of Canadian Productivity Growth

Simon van Norden[1]

Service de l'enseignement de la finance, HEC Montréal, 3000 Chemin de la Côte Sainte Catherine, Montreal, Québec, CANADA H3T 2A7

Abstract

For more than a decade, debates over the impact of new information technologies on trend productivity growth rates have played a key role in the formulation of monetary policy in many countries, including the United States and Canada. However, the question of whether the *trend* growth rate of aggregate productivity has changed *significantly* is rarely examined formally. This paper examines the latest aggregate labour productivity data for Canada using a new testing approach specifically designed to detect recent changes in trends. In addition to showing the strength of the evidence for changes in long-run trends, it considers the effect that data revision and changing sample period has had on inference about structural changes. In an appendix, it investigates how large such changes must be before they can be detected and to what degree detection tends to lag the structural change.

Evidence of a decline in the trend rate of labour productivity growth in Canada since 1990 is mixed. In particular, conclusions vary considerably from year to year as data are revised and as the accumulation of observations after purported breaks changes initial perceptions. The instability of test results suggest that policymakers need to use extreme caution in interpreting claims of changes in labour productivity trends and highlight the uncertainty that they face.

Key words: Productivity growth, detrending, breakpoints, structural change, data revision

Email address: simon.van-norden@hec.ca (Simon van Norden)

[1] HEC Montréal, CIRANO and CIREQ

Preprint submitted to North American Journal of Economics and Finance July 15, 2010

1. Introduction

Trends in productivity growth play a key role in the formulation of public policy. They are important factors in determining long-run economic growth and therefore play central roles in the management of public pension systems and government debt. They are an essential component in defining measures of economic slack and have therefore played a key role in the formulation of monetary policy, particularly in the United States and Canada.[2] International differences in such trends in turn have profound influences on the balance of world saving and investment. Not surprisingly, the possibility of a persistent change in aggregate productivity growth casts a long shadow over many of the most important economic policy debates. For all these reasons, great effort is devoted to accurate productivity measurement and to the analysis of sources of productivity growth.[3]

Surprisingly, however, the question of whether the *trend* growth rate of aggregate productivity has changed *significantly* is rarely examined formally.[4] The answer to this question is frequently ambiguous, even in the much-studied context of recent trends in U.S. productivity growth. For example, Gordon (2003) concludes

> Productivity growth experienced a second acceleration in 2000-03 following the initial productivity revival of 1995-2000. [*p. 272*]

a view shared by Bailey (2003).[5] Blinder and Yellen (2001) present

[2] For example, the May 4th 2004 minutes of the FOMC note the FRB staff's opinion that *"...that the remaining slack in resource utilization and strong productivity growth would keep core inflation at a low level over the forecast period."* The same minutes also summarize the committee's view that *"... a range of factors was continuing to restrain inflation, including slack in resource utilization, strong productivity gains ...."*

[3] Many different measures and definitions of productivity play a role in this research and policy debate. The most common are labour productivity and total-factor productivity (which considers aggregate inputs of both labour and capital.) This paper examines the behaviour of labour productivity (specifically, real output per employee.)

[4] This puzzling blindspot affects even the most important works in this field. For example, Gordon (2003) provides a 73-page analysis of the "explosion" in US productivity growth, but remarks by Durlauf and Sims [p. 293] lament the absence of any analysis of the uncertainty surrounding his estimates.

[5] "...it does now seem clear that trend labor productivity growth picked up substantially in the 1990s, and the most recent data suggest that there may even have been a further acceleration in the past two years. [*p. 282*]"

evidence for a change at the end of 1995, and Hansen (2001) finds that formal statistical analysis supports this timing.[6] However, Fair (2004) argues that late 1995 is simply a cyclical trough in productivity growth and that cyclically adjusted productivity growth shows very little improvement in the 1995-1999 period. Maury and Pluyard (2004) find weak evidence of trend break in US hourly productivity in 1995Q3, but none after the early 1980s when using per capita GDP.

We argue that at least two potential problems complicate the analysis in this literature. First, productivity data are revised over time, with the revisions often causing non-trivial changes in measured growth rates. None of the above papers investigates this problem in a systematic way. Second, very few papers perform statistical tests for changes in productivity growth trends, and some of those that do (Hansen (2001), Maury and Pluyard (2004)) use methods that are known to be unreliable close to the end of sample.

This paper examines the latest aggregate labour productivity data for Canada using a new testing approach specifically designed to detect recent changes in trends. Using real-time data, it then considers the impact of data revision on the detection of trend breaks. In an appendix, we also consider how large changes must be before they can be detected and to what degree detection tends to lag the structural change.

2. Literature Review

Most of the literature on trends in aggregate productivity growth relies on informal methods to characterize trends. On the basis of such analysis, it appears that profession opinion shifted around 2000 in favor of a improvement in US trend productivity growth which occurred less than five years before. For example, Stiroh (1999) looks at data up to 1998Q3 and concludes

> "... it is still too early to tell if a twenty-five-year trend of relatively slow productivity growth has been reversed."

while Jorgenson et al. (2002) write

[6] (Blinder and Yellen, 2001, p. 61, Table 8.1) and Hansen (2001, p. 123). Note that Hansen finds no evidence of a trend break in the early 1970s.

> "... that a consensus has emerged about trend rates of growth for output and labor productivity.... that the U.S. productivity revival is likely to remain intact for the intermediate future. [p.12]"

Blinder and Yellen (2001) present some informal evidence for a change at the end of 1995, but note that

> "As late as 1998...productivity growth had risen somewhat, but normal statistical tests attributed this rise to the cyclical upturn; there was no real evidence of any change in the underlying trend. [*p. 59*]"

Similarly, Gordon (1998) notes

> "...the failure of measured productivity growth to accelerate significantly in the 1990s. *[p. 299]*"

Writing in mid-year, Gordon (2000) notes

> "There is no dispute that [US] productivity has revived....*[p.49]*",

but then goes on to argue that the improvement is narrowly confined to computer manufacturing and that part of the apparent improvement is due to cyclical influences rather than a change in trend. However Baily and Lawrence (2001) write in early 2001

> "...there has been a substantial increase structural acceleration of total factor productivity (TFP) outside of the computer sector.*[p. 308]*"

and Gordon (2003) (as cited above) concludes that trend growth improved in the later part of the 1990s and then improved further in 2000-03.[7] Reviewing this literature, Edge et al. (2004) conclude that

[7] "These results refute my earlier position that actual productivity was above trend in the late 1990s; that position was defensible based on data at the time, but the soaring growth of actual productivity in 2002-03 now pulls the estimated trend lines well above actual values for the late 1990s. [p. 223]"

> "...the productivity acceleration in the late 1990s is now widely viewed as starting in the middle of the decade. ...[H]owever, estimates of long-run productivity growth by economists and professional forecasters changed little until 1999 and shot up dramatically in 2000."

The methods used to distinguish "trends" from other movements in productivity in the above papers are usually ad hoc; typical methods include the comparison of growth rates for non-overlapping five-year periods or the use of other arbitrary smoothing techniques (e.g. Gordon's use of the Hodrick-Prescott filter.) Lacking any formal statistical framework, such studies are unable to assess whether the changes in estimated trend rates are statistically significant, and therefore whether they are *reliable* indicators of structural economic changes. Exceptions to this approach are few and noteworthy; Borio et al. (2003) characterize this literature as showing that

> "The time series evidence of a structural productivity break was suggestive but not statistically significant until the late 1990s...*[p. 17, footnote 33.]"*

Filardo (1995) uses data through 1995Q2 to critically examine the claim of an increase in productivity growth in the early 1990s. While he finds evidence of a structural break around 1990Q4 for some tests, he concludes that more reliable tests find no significant evidence of a break. However, Hansen (2001) uses monthly data through April 2001 and finds significance evidence of a shift in the mid-1990s, with some series indicating a break around 1994 and others placing it in 1997. Maury and Pluyard (2004) use data ending in 2002Q4 and conclude that there is some evidence of a structural break in trend growth around 1995Q3 when using output per person-hour but not when using output per capita data. Erber and Fritsche (2005) use data ending in 2003Q4 and find breaks in the trend in 1994Q3 as well as 1998Q3. All four of these papers test for breaks by searching over all possible break dates and comparing the maximum F-statistic to a nonstandard distribution under the null. All also allow for multiple structural breaks using the iterative approach of Bai and Perron (1998).[8]

[8] Kahn and Rich (2007) use a very different approach to testing for changes in productivity growth trends. They estimate a multi-variate regime-switching factor model in

A more formal state-space analysis of productivity growth trends is provided by Benati (2007), who examines evidence for the US, the Eurozone and a selection of other economies (not including Canada.) He begins by noting that the above Bai and Perron tests find "surprisingly little" evidence of changes in productivity growth trends. He notes that this would be consistent with models in which the trend growth rate varies gradually over time rather than exhibiting sharp breaks, and proceeds to a careful estimation and examination of such models. However, he avoids formal tests of the null hypothesis of a constant trend growth rate in these models and finds that confidence intervals for the trend growth rate for the US are quite wide; the 90% interval at the end of sample has a width of 1.6% (in annual rates.)[9]

Filardo (1995) performs bootstrap simulations to determine the distribution of his F-statistic under the null hypothesis of no structural change, while Hansen (2001) and Maury and Pluyard (2004) use asymptotic critical values provided by Andrews (1993). However, it is well known that the test statistic converges only very slowly to its asymptotic distribution as the

which several variables (including productivity) switch between high and low trend growth states. They argue that their multivariate framework provides a more powerful test than the univariate tests used elsewhere. Their estimates date the break in trend in 1997 and the authors note that their model would have detected the regime switch within about six quarters of when it occurred.

Updates of their results published by the *Federal Reserve Bank of New York* show that their model's probabilities undergo important revisions for many years after first estimates are available. For example, the probability of being in the low growth state in 2006Q1 was estimated to be well under 10% in Sept. 2006, but well over 95% by Dec 2008, only to be revised downwards to under 60% 15 months later.

Their results rely *inter alia* on the strong assumption that any change in productivity growth trends after the 1970s and 80s must imply a "return" to patterns of the 1960s. Their implimentation of the regime-switching structure also precludes statistical tests for a structural break in the 1990s which are comparable to those discussed above. For example, although the authors show that "unsmoothed" estimates of the probability of being in the high-growth-trend regime exceed 95% by 1998Q4 [p. 17], this calculation ignores the effects of parameter uncertainty on regime inference; tests used in the abovementioned literature take account of this uncertainty.

[9] Benati provides graphs showing median-unbiased estimates of the trend growth rate for many series along with their 90% confidence intervals. Those for the US [p. 2864] are striking in that there is no hint of signficant evidence of a change in trend productivity growth; the minimum of the upper 90% bound is always above the maximum of the lower 90% bound. This is also typical of the results he show for other economies.

search for potential breakpoints comes "close" to the extremes of the data sample. For that reason, most applications of the this testing approach restrict the breakpoint not to lie within the first 15% or the last 15% of their observations. This is an awkward restriction when the object of interest is *current* trends in productivity growth. Fair (2004) uses a new test statistic proposed by Andrews (2003) which allows for breakpoint tests close to the end of the sample; however, while he applies the test to a broad range of macroeconomic series, he reports no test results for productivity series. We discuss the properties of this test in greater detail, below.

None of the papers discussed above consider the problems caused by the ongoing revision of data (except insofar as to reconcile an article's conclusions with previously published conclusions by the same author.) In contrast, Edge et al. (2004) carefully consider the problem of estimating trend productivity growth with real-time data.[10] They show that with such data, a heuristic linear updating rule can produce estimates of productivity growth which resemble historical estimates produced by the Council of Economic Advisors and other analysts. They also discuss the importance of data revisions in 1998 and late 1999 to changes in estimated trend productivity growth. However, this paper also stops short of providing a framework for formal statistical inference and uses a stylized state-space model only as a device to justify their use of a simple updating rule.

In contrast to the robust literature examining US data, papers examining Canadian data are far fewer and have tended to focus on identifying sources and causes of productivity growth rather than estimate its trends. Leading articles on aggregate productivity, such as Sharpe (2004), Macklem (2003) or Crawford (2002) discuss "trends" by performing growth-accounting of recent productivity data. They provide no distinction between trends and cycles in productivity beyond looking at changes over multi-year periods. Robidoux and Wong (2003) use the Hodrick-Prescott filter to measure trend growth. Despite their discussions of future trends in productivity and acknowledgement of the uncertainty surrounding such projections, none attempt to quantify this uncertainty in a statistical framework and none test for breaks in trend. Crawford (2002), Macklem (2003) and Sharpe (2004) contrast productivity developments in Canada against those in the US, but stop short of

[10] Kahn and Rich (2007) also provide some limited analysis of real-time data with vintages starting in the mid-90s.

considering whether there is any significant difference in their trend growth rates.

Reliable evidence of improvements in productivity growth trends requires reliable testing procedures. Below, we briefly review tests proposed by Andrews (1993) and Andrews (2003). Simulation evidence on their reliability is presented in an appendix. We then apply the latter test to original vintage data for Canada to replicate what researchers would have been to conclude about apparent changes in productivity growth trends and we examine the degree to which such conclusions are consistent over time. We also isolate the importance of data revisions in accounting for these results. Before that, however, we review how we choose to measure Canadian productivity.

3. Data

The data consist of seasonally-adjusted quarterly observations for Canada on real output at market prices per person employed. These in turn are based on quarterly figures for seasonally-adjusted real output at market prices from 1976Q1 onwards as reported by Statistics Canada.[11] The real output figures are then simply divided by the corresponding vintage figures for total employment seasonally-adjusted.[12]

3.1. Data Revision

In addition to analyzing the data as they are currently reported by Statistics Canada, we also examine "original vintage" series as they were published every quarter from 1976Q2 to 2009Q2. The "Final" series is the 2009Q2 vintage, whose last observation is for 2009Q1.

[11] Statistics Canada series *v1992067.* Note that prior to 1986 figures are for GNE rather than GDP; figures from the 2003Q3 vintage onwards are chain-weighted. Vintages up to the end of 2005 were provided by the Bank of Canada from their vintage model databases and from Statistics Canada reports. Figures thereafter were manually updated from Statistic Canada's *Canadian Economic Observer* (figures are taken from the first issue of each quarter.) Historical figures for the 2006-2009 vintages were taken from the most recent vintage in CANSIM; no benchmark revisions took place during this period.

[12] Statistics Canada series *v2062811.* Vintages up to the end of 2002 were provided by the Bank of Canada from Statistics Canada reports. Figures thereafter were manually updated from Statistic Canada's *Canadian Economic Observer;* the figures used are for employment in mid-quarter. Historical figures for the 2005-2009 vintages were taken from the most recent vintage in CANSIM; no benchmark revisions took place during this period.

Employment figures are regularly revised at the start of each year as the seasonal adjustment factors are updated; these revisions usually apply to last three years of data. Revisions are also carried out every five years as the surveys are reweighted to reflect information from the most recent census, with results from the July 2001 census incorporated in benchmark revisions at the start of 2005. Real output is revised when subsequent quarters for the same year are published and annually at the start of the year when the seasonal adjustment factors are published; the latter usually affects the last four years of published data. Historical revisions are also carried out once per decade.

3.2. Measuring Productivity

It would be desirable to examine figures covering only the business sector in order to avoid valuation problems in the non-business sectors. However, business-sector figures are available only from 1987Q1 onwards; this was judged to be insufficiently informative about trends prior to the early 1990s to be of use.[13] Quarterly total factor productivity estimates from the Bank of Canada QPM Simplex Database were available from 1966Q1 onwards.[14] However, this series is dominated by strong cyclical fluctuations. Given the large revisions of estimated cyclical factors at the end of sample, one would expect that a series with less pronounced cyclical factors should give more reliable estimates of recent trend growth rates Also, these estimates are not used outside the Bank of Canada; Statistics Canada has published only annual estimates of multi-factor productivity. Updates of the Statistics Canada series have been irregular and revisions frequent, as shown in Table 1.

3.3. Canadian Labour Productivity

The output per employee (OPE) data are summarized in Figures 1 and 2, and Table 3.3. Figure 1 shows the natural logarithm of labour productivity in the upper panel and its first difference in the lower panel. Both show the most recent (2009Q2) vintage estimates. We see that productivity is clearly cyclical with lower growth during the recessions in the early 1980s, 1990s, 2001 and the end of 2008. There are many periods in which productivity

[13] Due to the influence of a deep recession in the early 90s, this series shows a roughly constant level of productivity until the early 90s and steady growth thereafter.

[14] The author would like to thank Wendy Chan of the Bank of Canada for providing the data.

Table 1: Publication of Canadian Productivity Data by Statistics Canada

Issue Year	Year Released	Publication #	Notes
1984	unknown	14-201	
1985-86	1988	15-204 XIE	
1987	1989	15-204 XIE	Important structural breaks.
1988	1990	15-204 XIE	Important sectoral breaks.
1989	1991	15-204 XIE	Introduces multivariate TFP
1990-91	1992	15-204 XIE	Important structural breaks
1991	1993	15-204 XIE	
1992	1994	15-204 XIE	
1993	1995	15-204 XIE	
1994	1996	15-204 XIE	
	1997-2000		No publications
1999	2001	15-204 XIE	Major data revisions
2000	2002	15-204 XIE	
2003	2004	15-003	
2003	2005		Data revisions

appears to decline, and the volatility of productivity growth appears to drop after the mid-1980s. There is also some suggestion that productivity may be tending to "level off" after 2000.

Figure 2 gives some idea of the relative importance of data revisions by showing the real-time estimates of productivity growth. This real-time estimate is simply the growth rate in each vintage over its final 1, 4, 8 or 16 quarters. Each of the four panels in Figure 2 shows one of these four real-time estimates alongside the corresponding revision in that growth rate, where the revision is simply the difference between the estimate from the "final" (2009Q2) vintage and the real-time estimate. For ease of comparison, all series are shown at annual rates. We see that revisions are roughly the same order of magnitude as the real-time estimates of productivity growth. The volatility of the revisions seems to reflect the volatility of the real-time estimates; both vary widely and rapidly when measured over a single quarter, but unsurprisingly both become smoother as we measure growth rates over longer intervals. It also appears from the graphs that the average revision has been negative; productivity estimates have tended to be revised downwards over time.

Table 2: Revisions in Growth of Log Canadian Output per Employee.

Statistic	**1Q Change**	**4Q Change**	**16Q Change**	**32Q Change**
Real-time Mean	0.81	0.78	0.79	0.78
Real-time Std. Err.	2.00	1.10	0.81	0.62
Revision Mean	-0.19	-0.22	-0.22	-0.27
Revision Std. Err.	2.00	0.78	0.51	0.41
ρ (Real-time,Final)	0.577	0.790	0.861	0.796
ρ (Revision$_t$,Revision$_{t-1}$)	-0.183	0.665	0.769	0.888

Notes:

Final data is 2009Q2 Vintage; Real-time Vintages are 1976Q2-2009Q2.
All series start 1976Q2.
All figures are for change in logs at annual rates x 100.

Table 3.3 largely confirms what is shown in Figure 2. The upper portion of the table shows the mean of the real-time estimates of annual labour productivity growth. Estimates over 1, 4, 8 or 16 quarters all show very similar mean growth rates of 0.80% annually. As we expect from Figure 2, the standard error of these growth rates declines sharply from 2.0% for quarter-to-quarter changes, to 1.0% for 4-quarter changes and to 0.6% for changes over 16 quarters. The Table confirms that the mean revision has been negative, with productivity tending to be revised downward by about 0.20% from the 0.80% real-time estimate. The standard deviations of the revisions fall somewhat faster with horizon, falling from 2.0% for quarter-to-quarter changes, to 0.8% for 4-quarter changes and to 0.4% for changes over 16 quarters. The fifth line in the Table shows the correlation between the real-time and the final estimates of the growth rate of productivity; this rises from just under 60% at the one-quarter horizon to about 80% at the 4 and 16-quarter horizons and just over 85% when measured over 8 quarters. The final line shows that while revisions to the 1Q growth rates have no persistence and are actually negatively serially correlated, their persistence rises with horizon as we would expect, with a first-order autocorrelation coefficient of just over 88% at 16Q.

Together, Figure 2 and Table 3.3 suggest that revisions may be a potentially serious problem for the estimation of trend productivity. The standard error of revisions varies from roughly 65% to 100% of that of real-time productivity growth estimates; although the ratio improves as we move from short-run productivity movements to longer-term trends, it remains far from negligible and the persistence of revisions increases. The mean revision also appears to be non-zero; while small relative to the mean growth rate (about 1/4), it will likely be larger in relation to potential changes in productivity growth trends. All this may therefore complicate the timely detection of shifts in trend growth rates. We therefore turn in the next section to the problem of detecting such shifts.

4. Methodology

As noted above in the literature review, most of the work which tests for statistically significant changes in productivity growth trends use mid-sample tests, including those of Andrews (1993) and Bai and Perron (1998), which allow for single or multiple breaks at unknown dates. Below, we first review the construction of such tests as well as the problems they encounter when searching for "recent" breaks in trend.[15] We then explain the basis for a new test statistic proposed by Andrews (2003) which allows for breakpoint tests close to the end of the sample and we discuss its properties.[16]

4.1. Mid-Sample Tests

Chow (1960) proposes F-tests for a one-time structural change in one or more estimated regression coefficients when the date of the break is known. In the case of a simple AR(1) model, the null hypothesis is

$$(1) \qquad y_t = \rho \cdot y_{t-1} + \varepsilon_t \text{ for } t = 1, \ldots, T \text{ and } \varepsilon_t \sim IN(0, \sigma)$$

and the alternative is

[15] Our discussion focuses on the Andrew's test for a single breakpoint whose date is unknown. The Bai and Perron test generalizes this to multiple breakpoints, but shares the same characteristics with respect to its application near the end of samples.

[16] Yet another approach to modeling time-varying productivity growth trends would be to use the state-space model applied by Benati (2007). We note, however, that Benati does not provide a statistical test for changes in trend growth rates. Instead, he advocates modeling trend growth as varying continuously. We also note that Benati imposes assumptions of homoscedasticity similar to the ones we require, below. Finally, Benati's approach does not appear to have very high power to detect shifts in growth rates. As noted above, the 90% confidence intervals he graphs are wide and provide no evidence of significant changes in productivity growth trends at any time for any of the countries studied.

(1′) $y_t = \rho_0 \cdot y_{t-1} + \varepsilon_t$ for $t = 1, \ldots, \tau$ and $y_t = \rho_1 \cdot y_{t-1} + \varepsilon_t$ for $t = \tau + 1, \ldots, T$, where $\rho_0 \neq \rho_1$.

Let $\widehat{\varepsilon}, \widehat{\varepsilon}_0$, and $\widehat{\varepsilon}_1$ be the OLS residuals for these three equations and S, S_0 and S_1 be the sums of their squared residuals. The Wald test statistic for a structural break at τ is then given by

(3) $W = T \cdot \frac{S - S_0 - S_1}{S_0 + S_1}$

Andrews (1993) considers the distribution of this and related statistics when the researcher searches over possible values of τ. He proposes the test statistic

(4) Sup $W = \max_{\tau} W$ where $\pi \cdot T \leq \tau \leq (1 - \pi) \cdot T$ and π is referred to as a "trim factor".

Andrews shows that this statistic converges to a nonstandard distribution under very general conditions and provides tabulated asymptotic critical values. He also shows that the test will generally have better asymptotic power than other stability tests (such as the CUSUM.) Andrews and Ploberger (1994) provide stronger optimality results for a closely related statistic,

(5) $$\text{Exp } W = \ln \left[\sum_{\tau} \exp \frac{W}{2} \right] + \ln \left(T \cdot (1 - 2\pi) \right)$$

Hansen (1997) provides formulas for approximate *p-values* for both the Sup W and the Exp W statistics; both statistics are used in Hansen (2001). Throughout this literature, the asymptotic theory requires that π is constant and bounded away from 0 as $T \to \infty$. We will therefore refer below to this test as a *mid-sample* test to distinguish it from the *end-of-sample* test introduced below. Andrews (1993) suggests using $\pi = 0.15$, and this choice is widespread in the applied literature.

4.2. End of Sample Tests

Andrews (2003) proposes a related approach for testing stability close to the end of the sample. Let ε_1^R be the subvector of $\widehat{\varepsilon}$ and X_1 be the set of regressors (in this case, just y_{t-1}) for $t = \tau+1, \ldots, T$, and let $\widehat{\sigma} = S'S/(T-1)$. Andrews' test statistic is then[17]

(6) $$\psi = \widehat{\varepsilon}_1^{R\prime} X_1 \left(X_1' X_1\right)^{-1} X_1' \varepsilon_1^R / \widehat{\sigma}^2$$

Unlike the Andrews (1993) approach, this statistic does not compare parameter estimates before and after the breakpoint. Instead, it compares full-sample estimates of the residual variance to the size of the (transformed) residuals near the end of sample. Large values of the latter relative to the former are evidence of a structural break. Tests of this form may thought of as "predictive failure" tests, or tests of a quite general hypothesis of structural stability. In the above AR(1) example, such a statistic would tend to reject the null hypothesis of no structural change even when $\rho_0 = \rho_1$ but σ increases after the breakpoint. We return to this point, below. The Andrews (2003) test also differs from the Andrews (1993) test in that the date of the breakpoint is treated as known.

The distribution of the test statistic under the null hypothesis is nonstandard; Andrews (2003) proposes a subsampling-based simulation approach to tabulate appropriate *p-values* in specific applications. This consists of calculating the test statistic for all possible samples of length $T - \tau$ over the period $t = 1, \ldots, \tau$ in order to estimate its distribution under the null hypothesis of stability.

Since the correct breakpoint for the Andrews (2003) test is not known, we must search over possible break dates, which presents a problem in maintaining the correct size of the test. Therefore, we also consider a generalization of the above ψ test statistic in which τ is unknown. The resulting test statistic is simply the maximum of the ψ statistic defined above over all values of $\tau | \pi \cdot T < \tau < T$. Critical values are similarly computed using the subsampling approach, which now requires calculating $\max\left(\psi\left(\tau\right)\right)$ for all possible samples of length $T - \tau$ and all values of τ. We note, however, that the Andrews (2003) test has reasonable power to reject the null hypothesis even when the date of the structural break is (slightly) misspecified. We therefore consider evidence from the $\max\left(\psi\right)$ test alongside that from the ψ test for a fixed breakpoint of $\pi \cdot T$.

[17] See Andrews (2003) for a general exposition. Note that when the number of regressors is greater than or equal to the number of post-break observations, the test statistic is instead calculated as $\psi = \widehat{\varepsilon}_1^{R\prime} \varepsilon_1^R / \widehat{\sigma}^2$.

4.3. Heteroscedasticity

Unlike the mid-sample test, the end-of-sample test relies on the *joint* null hypothesis of no structural break and homoscedasticity. As can been seen from the above equation, an increase in the variance of the residuals at the end of the sample will tend to increase the size of the test, while a decrease in their variance will decrease its power. Using the Andrews (1993) test on *squared* productivity growth finds significant evidence of a decrease in the volatility of productivity growth with the most likely date for the break being 1987Q2. Assuming that this break would have been apparent by 1990, we therefore adjust all our data series for this shift prior to applying the Andrews (2003) test. Specifically, let $\widehat{\sigma}_1$be the estimated standard deviation of productivity growth about its sample mean up to 1987Q2 and $\widehat{\sigma}_2$ be the standard deviation thereafter. We test the transformed series $\widetilde{y}_t$ where

(7) $\widetilde{y}_t = y_t$ up to 1987Q2

$\widetilde{y}_t = \widehat{\mu} + (y_t - \widehat{\mu}) \cdot \widehat{\sigma}_1 / \widehat{\sigma}_2$ where $\widehat{\mu}$ is the sample mean after 1987Q2.

$\{\widehat{\mu}, \widehat{\sigma}_1, \widehat{\sigma}_2\}$ are estimated only on the vintage series that they are used to transform.

5. Recent Breaks in Trend Productivity Growth

5.1. Known Breakpoint Tests

We now consider tests for recent breaks in Canadian labour productivity growth trends. We transform the data described above using (7) and perform both the Andrews (2003) test as well as the generalization we described above for the case of unknown breakpoints. Specifically, we estimate

(8) $y_t = \alpha + \rho \cdot y_{t-1} + \varepsilon_t$ for $t = 1, \ldots, T$

and

(8′) $y_t = \alpha_0 + \rho_0 \cdot y_{t-1} + \varepsilon_t$ for $t = 1, \ldots, \tau$ and $y_t = \alpha_1 + \rho_1 \cdot y_{t-1} + \varepsilon_t$ for $t = \tau + 1, \ldots, T$

and test the joint null hypothesis that $\rho_0 = \rho_1$ and $\alpha_0 = \alpha_1$ using the ψ statistic given in (6).[18] We repeat the test for every quarterly vintage of our

[18] Note that the unconditional mean of an AR(1) series is given by $\alpha / (1 - \rho)$. We therefore focus on joint tests for the stability of both α and ρ rather than on tests for just α.The tests were repeated for an AR(2) specification with tests of the joint null hypothesis of stability in all three coefficients. The results were essentially the same to the results for the AR(1) case presented below.

productivity series from 1990Q1 to 2009Q2. For each of these vintages, we then repeat the test for all values of $\tau | 0.85 \cdot T \leq \tau < T$. This enables us to both see whether there is significant evidence of a break in the trend growth rate of productivity, and when such evidence would have become apparent.

The *p-values* of the resulting tests are shown in Figure 3, where T is shown on the horizontal axis and τ on the vertical. The legend shows the color assigned to *p-values* in various ranges, with only dark blue areas indicating significant evidence of changes in productivity growth trends. Consistent detection of the same breakdate in subsequent vintages should appear as a blue area that extends horizontally along a given value of τ. This area will start at the 45-degree line (i.e. $\tau = T$) if the break is detected the first time it is tested. However, if more observations are required to detect a break, then the blue area would start to the right and below the $\tau = T$ line. Vertical features are associated with data revisions; vertical contours (i.e. a change from one colour to another) indicate that the use of different vintages causes important changes in the perceived probability of a trend break.

There are several noteworthy features in Figure 3.

1. There is some, but not much, dark blue in the graph. Significant evidence of breaks in productivity growth trends can be found and is mostly confined to small number of distinct episodes.
2. Most of the evidence of structural breaks is found during recessions; low *p-values* are found in data vintages from the start of 1991, 2001 and the end of 2008. Others include a very brief period in 1996 and a longer period in 2006.
3. Evidence of structural breaks tends to vanish abruptly as new data become available. Evidence of a shift around the 2001 vintage disappears at the start of 2002, while that around the 2006 vintage vanishes within a year. A possible exception to this may be the evidence which appeared in the late 2008 vintages, but it is too soon to draw a reliable conclusion. These vertical features suggest that changes from one vintage to the next are sometimes important in creating and destroying evidence of breaks in productivity growth trends.
4. Some of the most suggestive evidence for a change in productivity growth trends comes from tests for a change in 1999Q1&2, with strong evidence that persisted until a May 2002 benchmark revision in GDP, whereupon all evidence vanished. There is also fairly persistence evidence of a break near the end of 2005 that can be found in most vintages

up to and including the most recent. As we can see from Figure 1, this corresponds to a downward shift in the trend growth rate.

The inconsistency of test results for the same break date on different data vintages is troubling as it underscores the uncertainty confronting forecasters and policymakers. One possible explanation for this inconsistency is that data revisions have significantly altered our perception of productivity growth trends. Another is that the addition of more observations after the purported break date should be very informative in understanding whether or not trend growth has changed. To better understand the role of these factors, we repeat the analysis shown in Figure 3, but now remove the effects of data revisions by using only the corresponding observations in the 2009Q2 data vintage. The results are shown in Figure 4. They differ from those in Figure 3 in at least two important respects.

1. There are more dark blue patches, particularly in the 1990s, providing more evidence of shifts around 1990 and 1994. However, both of these were periods when productivity growth initially stalled before resuming shortly thereafter (see Figure 1) suggesting that early rejections of the null hypothesis of no structural breaks may have been spurious.
2. There are more horizontal features in the graph (including most even-numbered years from 1990 to 2000), suggesting that results are now more consistent across time. Despite this, many vertical features are still evident (such as in 1994, 2001 and the end of 1998.) This suggests that both data revision and longer post-break data samples have played important roles.

To better understand this last point, we turn to Figure 5, which simply graphs the difference in *p-values* between Figures 3 and 4. These changes are entirely due to the effects of data revision. Not surprisingly, we find several vertical features in the graph. Of particular interest are the darker blue patches and red patches. The former indicate areas where original data vintages showed very low (significant) *p-values* while the revised data show every high (insignificant) *p-values*. The indicate two episodes, the first in 2001 and the second at the end of 2006, where initial estimates produced spurious evidence of changes in productivity growth trends that vanished when data were revised.[19] The opposite occurred more frequently (partic-

[19] Note that the blue patches do not extend to the top of the coloured diagonal band.

ularly in the 1990s); data revision produced strong evidence of changes in productivity growth where none was found previously. This implies that, despite the power of the end-of-sample test to detect relatively small changes in growth trends, evidence may only become apparent after a delay of a few years due to data revisions.

To put this into perspective, it may be useful to remember the size of the changes in productivity growth underlying these results. As noted in the appendix, labour productivity growth in our data sample averaged 1.4% annually. Figure 2 shows that the 16Q average growth rate varied between 1.8% and 0% in initial releases over the period shown in the graph, while revisions could raised these average by up to 0.5% or lower them by up to 0.8%. The simulations performed in the Appendix and summarized in Figure A3 imply that these tests should be able to detect persistent changes of 0.7% within 5 year about half the time, and persistent changes of 1.4% within 2 years slightly more than half of the time.

5.2. Unknown Breakpoint Tests

A reasonable criticism of the test results presented so far in this section is that they may suffer from repeated test bias; with literally hundreds of breakpoint tests performed in each graph, it is not surprising *a priori* to find that some test statistics exceed the nominal 5% or 1% critical values. We can mitigate this problem (but not eliminate it) by using tests for unknown breakpoints, which provide properly sized tests for each data vintage. As described above, we use both the $\max(\psi)$ test alongside that from the ψ test for a fixed breakpoint of $\pi \cdot T$. *p-values* are again shown in Figure 6, with those for the ψ test in green and those for the $\max(\psi)$ test in red. The blue line in the upper half of the Figure (graphed on the right-hand scale) shows the estimated break date for the $\max(\psi)$ test.[20] The horizontal axis shows

This means that the data revisions did not affect tests on the most recent few quarters. Rather, they changed conclusions about the possiblity of productivity growth changes several years earlier.

[20] That is, it gives the value of τ which maximizes ψ. Because breaks for this test are constrained to be near the end of sample, the estimated breakpoint trends upwards over time. These estimated breakpoints should be interpretted together with the plotted p-values for the $\max(\psi)$ test. For example, the estimated break date is consistently estimated to be in 2001Q4 when viewed with many data vintages from 2002 and 2003. However, the p-values from the $\max(\psi)$ test (red line) are consistently near 1 throughout this period, indicating that the evidence is not significant.

the value of T for each vintage tested.

We see from the figure that there are only two periods in which there appears to be significant evidence of changes in productivity growth trends; the start of 2001, when both tests briefly provide strong evidence against the null hypothesis of no break, and (in the case of the $\max(\psi)$ test only) at the end of 2008. Evidence from the latter test suggests a break in 1997 during the first episode and in 2008 during the latter.[21] Both are consistent with the results in Figure 3, and correspond to the most consistent and significant evidence of trend breaks. The vertical feature in Figure 3 around 2006 also corresponds to the spike in Figure 6 at the same period, which produces a *p-value* of just over 8%. It therefore appears that the most important features of tests shown in Figure 3 are not altered by the correction for repeated test bias

6. Conclusions

The results presented above support two main conclusions. The first is that there appears to be evidence of a slowdown in productivity trend growth in Canada that took place around 1997. There is also recent evidence of a further slowdown that coincides with the 2008 recession.

The second conclusion, which may be of more general applicability, is that such statistical evidence of structural changes is frequently fragile. We have shown that data revision may overturn strong statistical evidence of structural changes, or create it where none has existed before. Furthermore, this does not appear to be the result of the refinement of preliminary estimates in the first quarter or two after their release; rather, data revisions may lead to re-evaluation of events several years after the fact. We have also seen that the data revision is not the sole factor responsible for the fragility of statistical test results. The accumulation of observations can also be very informative about the validity of perceived changes in productivity trends. Even in the absence of data revision, therefore, judgements about productivity growth trends may be revised for many years after the fact.

[21] Althought the evidence in the former case is fleeting, vanishing after only quarters, this simply reflects the fact that with $\pi = 0.15$, the purported break date moves out of the test window two quarters after the break becomes statistically significant. Mid-sample tests on longer vintages also produce signficance evidence of a slight decline in productivity growth around this time.

The instability of test results suggest that policymakers need to use extreme caution in interpreting claims of changes in productivity trends and highlight the uncertainty that they face. Improving the reliability of existing statistical methods remains an important task for future research.

Acknowledgement 1. *The author would like to thank Wendy Chan, Mary McGuire and Marty Olson of the Bank of Canada for their assistance in making vintage data available and Clare van Norden for her help in updating the data set. Thanks also to Bruce Hansen for making code available from his 2001 Journal of Economic Perspectives article, and to seminar participants at Rutgers University, the Bank of Canada and the CEF Meetings in Washington D.C. for their comments. Finally, thanks to Robert Amano at the Bank of Canada for encouraging me to undertake this project and to Shaun Vahey for encouraging me to update it. Portions of this research have been financed by the Bank of Canada and by the New Economic Initiative of the Social Sciences and Humanities Research Council of Canada. The opinions expressed herein are the author's and should not be confused with those of any of the above. The author retains sole responsibility for all the errors, defects and shortcomings of this work.*

References

Andrews, Donald W. K. (1993) "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, 61(4), 821-56.

Andrews, Donald W. K. (2003) "End-of-Sample Instability Tests." *Econometrica*, 71(6), 1661-94.

Andrews, Donald W. K. and Werner Ploberger (1994) "Optimal Test When a Nuisance Parameter is Present Only Under the Alternative" *Econometrica*, 62:6, 1383-414.

Bai, Jushan and Pierre Perron (1998) "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica*, 66(1), 47-78.

Bailey, Martin Neil, (2003) "Comments and Discussion on 'Exploding Productivity Growth: Context, Causes and Implications' by Robert Gordon" *Brookings Papers on Economic Activity*, 2:2003, 280-287.

Baily, Martin Neil and Robert Z. Lawrence (2001) "Do We Have a New E-Conomy?" *The American Economic Review*, 19(2), 308-312.

Benati, Luca (2007) "Drifts and Breaks in Labour Productivity" *Journal of Economic Dynamics and Control*, 31, 2847-2877.

Blinder, Alan S. and Janet L. Yellen, (2001) *The Fabulous Decade: Macroeconomic Lessons from the 1990s*, The Century Foundation Press, New York.

Borio, Claudio, William English and Andrew Filardo (2003) "A tale of two perspectives: old or new challenges for monetary policy?" *BIS Papers*, 19.

Chow, Gregory C. (1960) "Tests of Equality Between Sets of Coefficients in Two Linear Regressions" *Econometrica*, 28:3, 591-605.

Crawford, Allan (2002) "Trends in Productivity Growth in Canada" *Bank of Canada Review*, Spring 2002, 19-32.

Edge, Rochelle M, Thomas Laubach and John C. Williams (2004) "Learning and Shifts in Long-Run Productivity Growth" *Federal Reserve Board Finance and Economics Discussion Series*, 2004-21.

Erber, Georg and Ulrich Fritsche (2005) "Estimating and Forecasting Aggregate Productivity Growth Trends in the US and Germany." *DIW Berlin, Discussion paper* 471.

Fair, Ray C. (2004) "Testing for a New Economy in the 1990s" *Business Economics*, January 2004, 43-53.

Filardo, Andrew J. (1995) "Has the Productivity Trend Steepened in the 1990s?" *Federal Reserve Bank of Kansas City Economic Review*, 1995Q4, 41-59.

Gordon, Robert J. (2000) "Does the 'New Economy' Measure up to the Great Inventions of the Past?" *The Journal of Economic Perspectives*, 14(4), 49-74.

Gordon, Robert J. (2003) "Exploding Productivity Growth: Context, Causes and Implications", *Brookings Papers on Economic Activity*, 2:2003, 207-279.

Gordon, Robert J. (1998) "Foundations of the Goldilocks Economy: Supply shocks and the Time-Varying NAIRU" *Brookings Papers on Economic Activity*, 2:1998, 297-346.

Hansen, Bruce E. (1997) "Approximate Asymptotic *P Values* for Structural-Change Tests" *Journal of Business and Economic Statistics*, 15:1, 60-67.

Hansen, Bruce E. (2001) "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity", *Journal of Economic Perspectives*, 15(4), 117-128.

Jorgenson, Dale W., Mun S. Ho and Kevin J. Stiroh (2002) "Projecting Productivity Growth: Lessons from the U.S. Growth Resurgence" *Federal Reserve Bank of Atlanta Economic Review*, III2002, 1-13.

Kahn, James A. and Robert W. Rich (2007) *Tracking the New Economy: Using Growth Theory to Detect Changes in Trend Productivity,* Journal of Monetary Economics 54, 1670-1701

Macklem, Tiff (2004) "Future Productivity Growth in Canada: comparing to the United States" *International Productivity Monitor*, 7(Fall), 50-57.

Maury, Tristan-Pierre and Bertand Pluyaud (2004) "The Breaks in Per Capita Productivity Trends in a Number of Industrial Countries." *Notes d'études et de recherche de la Banque de France*, 111.

Robidoux, Benôit and Bing-Sun Wong (2003) "Has Trend Productivity Growth Increased in Canada?" *International Productivity Monitor*, 6(Spring), 47-55.

Sharpe, Andrew (2004) "Recent Productivity Developments in Canada and the United States: Productivity Growth Deceleration versus Acceleration." *International Productivity Monitor*, 8(Spring), 16-26.

Stiroh, Kevin (1999), "Is There a New Economy?" *Challenge*, 42(4), 82-101.

A. Appendix

This appendix presents evidence on the size and power of both end-of-sample and mid-sample breakpoint tests.

A.1. Mid-sample Tests

We begin by considering the evidence for a trend break in US productivity growth presented by Hansen (2001). As mentioned above, Andrews' mid-sample test for structural breaks requires that the breakpoint tested is "not too close" to the end of the sample in order to be asymptotically valid. Unlike most studies, Hansen (2001) tests for break by using the mid-sample test to within 5% of the end of the sample.[22] To assess the reliability of the Andrews test in this setting, we conduct a closely related simulation experiment. Hansen (2001) uses the monthly growth rate of the ratio of the US Industrial Production Index for manufacturing/durables to average weekly labor hours from February 1947 to March 2001 to estimate the model.

$q_t = \alpha + \rho \cdot q_{t-1} + \varepsilon_t$

and test for the joint significance of a structural break in α and ρ.[23] He finds that Andrews' mid-sample test has an approximate p–value of 0.0016 and concludes "We are therefore quite confident that this time series has a structural break.[p. 120]"

To investigate the reliability of this conclusion, we apply the same test to data simulated under the null hypothesis of no structural break. Specifically,

1. We estimate $(\widehat{\alpha}, \widehat{\rho})$ in the above equation using OLS on the full sample period (652 monthly observations covering roughly 55 years from 1947 onwards.)
2. We draw a random value of t and set $q_0^i = q_t$.
3. From the OLS residuals $\widehat{\varepsilon}$, we randomly draw (with replacement) T observations ε^i.
4. Using $(\widehat{\alpha}, \widehat{\rho})$, q_0^i and ε^i, we simulate a new series q^i.
5. For a given trim factor π, we use the simulated series to calculate the Andrews mid-sample test statistics Sup W^i and Exp W^i.
6. We use Hansen (1997) to convert the statistics into purported *p-values* and store the results.
7. We repeat steps 2 through 6 10,000 times

If the tests are correctly sized, then the resulting *p-values* should have a uniform distribution. Table 3 presents the frequency with which low *p-value*s

[22] Hansen (2001) uses 650 monthly observations; a 5% trim factor implies breakpoints may be as little as 33 months from the end of the sample.

[23] Hansen (2001) also tests for and finds no evidence of a structural break in σ.

Table 3: Joint *SupW* Tests for Parameter Stability

Trim	Fraction of *p-values*		
$\boldsymbol{\pi}$	$<$ **10**%	$<$ **5**%	$<$ **1**%
0.20	0.1127	0.0608	0.0123
0.15	0.1173	0.0594	0.0136
0.10	0.1260	0.0691	0.0140
0.05	0.1526	0.0858	0.0229

Note: Figures shown above are asymptotic *p-values* under H_0

are observed, and the overall distribution of *p-value*s is summarized in Figure A1. Following Hansen (2001), we also present only results for the joint test of stability in $(\widehat{\alpha}, \widehat{\rho})$.[24]

Results for the Sup W^i and Exp W^i tests were very similar with the latter performing slightly better; for brevity we therefore omit the former from the table. The results in Table 3 show that the Sup W^i test is well-sized for trim factors of 15 and 20%, with rejections at the 5% significance level occurring only about 6% of the time under the null. As the trim factor is reduced beyond that point, the test becomes increasingly liberal. In the worst case, with a 5% trim, the null is rejected about 50% more frequently than it should be at the 10% significance level and more than twice as often at the 1% level.

The two upper panels of Figure A1 summarizes these results in a *p-value* plot. Each panel displays the difference between cumulative distribution of the simulated *p-value*s and their theoretical distribution under the null (i.e. uniform on (0,1).) In the absence of size distortion, therefore, these lines should line along the x-axis. We restrict our attention to the (0,0.1) interval of the purported *p-value*s as this is the range most relevant for hypothesis testing. Results for the Sup W^i and Exp W^i tests are shown in the left and right panels, respectively.

Figure A1 again shows that the 20% and 15% trim factors in this case produce only modest size distortion; for example, a 5% nominal *p-value* for

[24] Results for the individual statistics were also examined. Tests for the constant were generally very well sized. Size distortion in the joint test generally reflected size distortion in the test on the autoregressive coefficient. Of course, instability in *either* implies instability in the mean growth rate.

the Exp W^i statistic produces a size distortion of roughly 1%, giving an actual size of roughly 5 + 1 = 6%. However, size distortion with a 10% trim is roughly double that of the 15%, and the distortion roughly doubles again in moving from the 10% to the 5% trim. The same 5% purported *p-value* for the Exp W^i statistic now produces an actual size of roughly 5 + 3.5 = 8.5% This corresponds well to the figures shown in the 5% column of Table 3.

These results confirm that while the mid-sample stability test appears to perform well with a trimming factor of 15%, testing closer to the end of the sample tends to produce too many spurious rejections of the null hypothesis of stability, with the severity of the problem increasingly sharply as the trim factor is reduced. This size distortion is not enough to undermine Hansen (2001)'s conclusion about a structural change in trend productivity growth, however. Using a 5% trim factor, his reported Sup test statistic of 20.2 produces a simulated *p-value* of 0.97% (versus the 0.16% asymptotic *p-value* he reports.) Accordingly, his confidence in the presence of a structural break in this series appears to be well-founded.

These simulation results pose a more general problem, however. The need to avoid smaller trim factors in order to have reliable interpretations of these test statistics prevents policy-makers from testing for recent breaks. This is not a trivial problem. For example, one cannot simply apply the same trimming factors to a shorter data sample, thereby effectively testing for breaks closer to the present time. The fundamental problem is one of a lack of sufficient observations after the break point to have the test statistic converge to its asymptotic distribution. To illustrate the problem, we repeated the above simulation reducing the sample size from 652 to 120 observations (i.e. 10 years of monthly data.) Results are summarized in the lower two panels of Table 3. A comparison of the vertical scales of the upper and lower panels in Figure 1 confirms that the degree of size distortion is now many times greater than before. In fact, if we compare results for π=0.05 in the longer sample (33 observations trimmed) to those for π=0.20 in the shorter sample (24 observations trimmed), we can see that their results are similar. For example, asymptotic *p-value*s smaller than 5% are observed roughly 9% of the time in both cases for both test statistics. [25]

[25] As we discuss below, replacing asymptotic p-values with simulated values is also not a panacea. While this corrects test size, low trim factors cause the test to have lower size-adjusted power.

A more reliable analysis of the recent past therefore requires a different approach, such as the use of the Andrews (2003) test.

A.2. End-of-Sample Tests

To understand the finite-sample performance of the end-of-sample test, we repeated the above experiment using the Andrews (2003) test. Specifically,

1. We estimate $(\widehat{\alpha}, \widehat{\rho})$ as before.[26]
2. We set the break date to be tested to correspond to the last break date covered in the mid-sample test with a trim factor of 15%.
3. We draw a random value of t and set $q_0^i = q_t$.
4. From the OLS residuals $\widehat{\varepsilon}$, we randomly draw (with replacement) T observations ε^i.
5. Using $(\widehat{\alpha}, \widehat{\rho})$, q_0^i and ε^i, we simulate a new series q^i.
6. Given the break date, we use the simulated series to calculate the Andrews end-of-sample test statistics Exp W^i.
7. We use the simulated series q^i to estimate the *p-value* of the test statistic using Andrew's parameter sub-sampling procedure and store the results.
8. We repeat steps 3 through 7 5,000 times.
9. We increment the break date to be tested, repeating steps 3 through 8 for each date, until we reach the end of the sample.

Again, if the tests are correctly sized, then the resulting *p-value*s should have a uniform distribution. Table 4 presents the frequency with which low *p-value*s are observed, and the overall distribution of *p-value*s is summarized in Figure A2.

The results show that, unlike the mid-sample tests, the end-of-sample test is most accurately sized at the end of the sample, and tends to become increasingly liberal as the breakpoint moves away from the sample end. Among other things, this presumably reflects the fact that sample available for bootstrapping the test statistic under the null shrinks as the breakpoint moves away from the sample's end, thereby reducing the precision of the simulation.

[26] To reduce the computational burden, we first converted the data from monthly to quarterly frequency by summing productivity for all months in each quarter.

Table 4: Joint S Tests for Parameter Stability

Periods from EOS	**10% p-value**	**5% p-value**	**1% p-value**
2	10.1%	5.1%	1.9%
4	9.9%	5.7%	1.7%
8	11.2%	6.2%	1.7%
16	11.3%	6.9%	2.3%
32	14.5%	9.0%	5.1%

Frequency of Parametrically Resampled *p-values* Under the Null

A.3. Power

While the simulations in the preceding section provide a better understanding of the tests' size properties, they give no indication of their power. However, policy makers need to know how large trend breaks in productivity must be before they are likely to be detected and how quickly breaks may be detected in order to better weigh their potential policy errors. It would therefore be useful to know whether breaks are likely to be recognized within a few quarters, or whether several years pass before enough evidence is available to reliably conclude that a change in trend has occurred.

To answer this question, we estimated the power of the end-of-sample test using a series of simulation experiments similar to those used above to establish its size. However, in doing so we found that the homoscedasticity. assumption used in the end-of-sample test might be unrealistic and have an important effect on the results. As shown in the lower panels of Figure 1, the volatility of quarterly productivity growth appears to be much lower after the 1980s than before, which would be consistent with other evidence of increased macroeconomic stability in the Canadian and other economies after the early 1980s.[27] For Canada, the annualized standard deviation of quarterly log productivity growth fell by just over half from 3.6% to 1.7%. As discussed above, such a decrease in the volatility of productivity should cause the end-of-sample test to lose power, as relatively larger breaks are now required to generate significant outliers. Given that the lower volatility appears to persist up to the present, the power of the end-of-sample tests in this newer environment should be more relevant for policy makers than their

[27] The Andrews (1993) test found evidence of a break in the variance of shocks to quarterly productivity growth in both countries that was significant at the 0.1% level.

performance using the average volatility over the available sample period.

To estimate the power of the end-of-sample test, we used an experimental design very similar to that used above to investigate test size, this time calibrating our experiment to the Canadian OPE data under the null hypothesis of no breaks. To correct for the apparent heteroscedasticity discussed above, OLS regression residuals before the break were rescaled so that their mean squared errors equalled that of the OLS residuals after the break. Bootstrapped residuals were then drawn with replacement from these rescaled residuals. To generate data under the alternative hypothesis of a break in trend, we now introduced a multiplicative constant k to the intercept term in the AR representation for the data. In all cases, $k = 1$ until the date for the break in the mean growth rate, after which it takes on a new constant value. This model was used together with the bootstrapped rescaled residuals to generate artificial data sets. End-of-sample break tests were then run on each of 5000 artificial data sets and estimated *p-value*s were tabulated. This was repeated for all possible break dates over the last 15% of the data sample. Tests were always correctly specified in the sense that the break date tested always corresponded to the true break date. (Misspecification of the break date would presumably lower the power.)

Figure A3 summarizes the results for the Canadian OPE data, where the annualized rate of productivity growth under the null of no breaks is 1.4% annually.[28] The four panels show the frequency with which the test produced *p-value*s lower than or equal to the value shown on the vertical axis while the horizontal axis indicates the number of periods after the break which are being tested. Blue values indicate that low *p-value*s were infrequent (i.e. low power), red values indicate frequent low *p-value*s (high power) while greens and yellows indicate intermediate results. For reference, the panel in the bottom right simulates the data under the null hypothesis of no breaks and provides evidence on the size of the test similar to that reported previously in Figure A2.

The four panels show that as the magnitude of the breaks increase, contour lines move down and the left. This means that the probability of detecting a break after a given number of periods increases with the size of the break. It also implies that the time required to detect a break with a given

[28] This 1.4% is higher than the sum of the mean real-time growth rate (0.6%) and the mean revision (0.5%) reported in Table 1 due to differences in the sample period.

probability level falls with break size. For example, the top right panel shows that after 5 years (20 quarters) the probability of detecting a break at the 5% significance level is about 90% for a doubling (from 1.4% to 2.8%) of the trend growth rate. The detection probability is roughly 40% for a 50% rise in the growth rate, and 100% for a tripling of productivity growth Put another way, the time required to detect an improvement in productivity growth at the 5% significance level with a probability of at least 50% is about 6 quarters in the case of a doubling of trend growth, about 5 years for a 50% rise and less than one year for a 200% improvement. These results imply that the end-of-sample test can be quite powerful for large enough changes; improvement of over 2% per year will be detected with high probability in a few quarters at conventional significance levels. However, smaller but still economically significant improvements (on the order of 0.5 to 1%) require substantially longer periods before convincing statistical evidence of a change is likely to be found.

Figure 1

Canadian Real Output per Person Employed

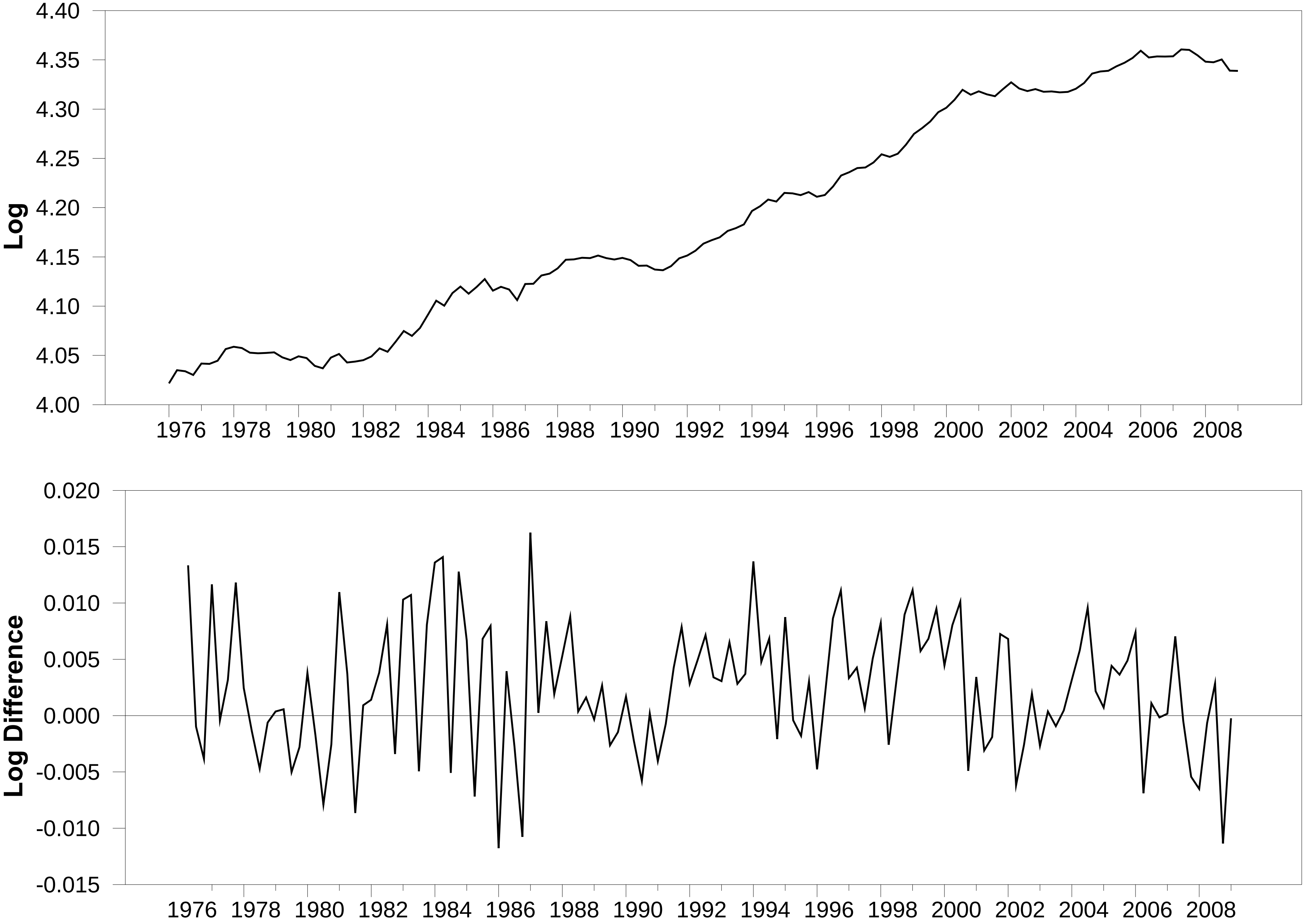
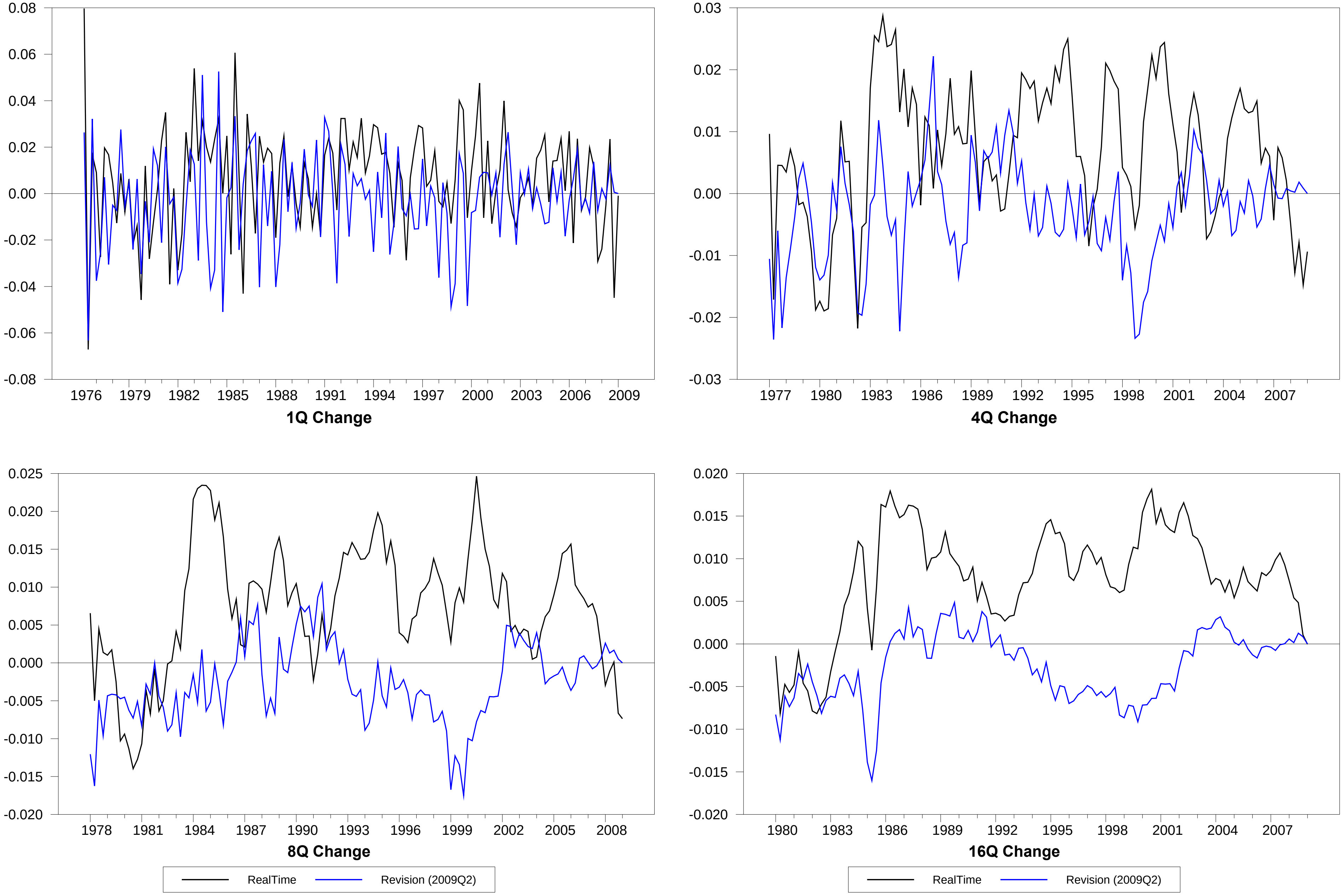


Figure 2

Productivity Growth Revisions



1Q Change

4Q Change

8Q Change

16Q Change

Figure 3 - End of Sample Breakpoints - RealTime

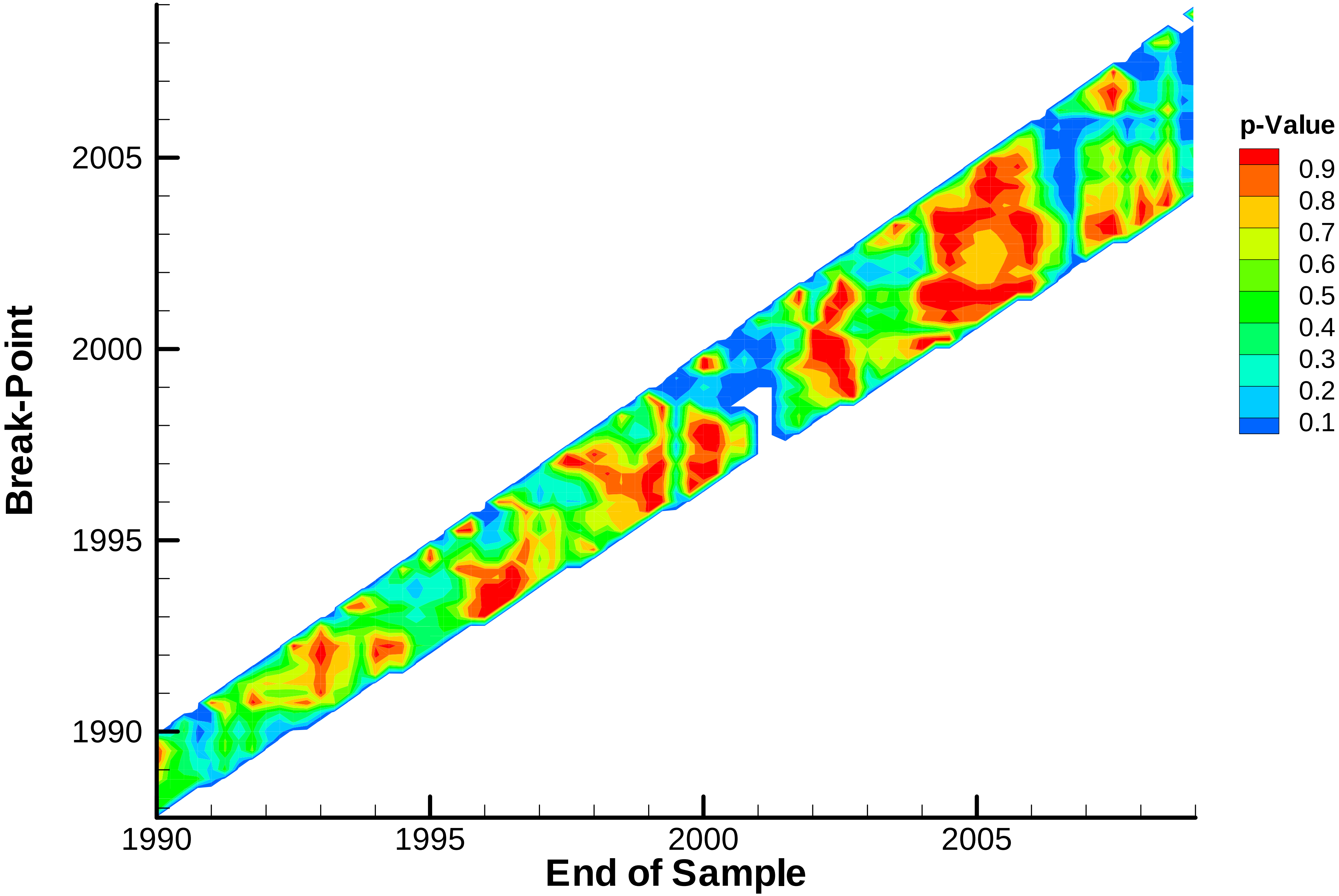


Figure 4 - End of Sample Breakpoints - QuasiReal

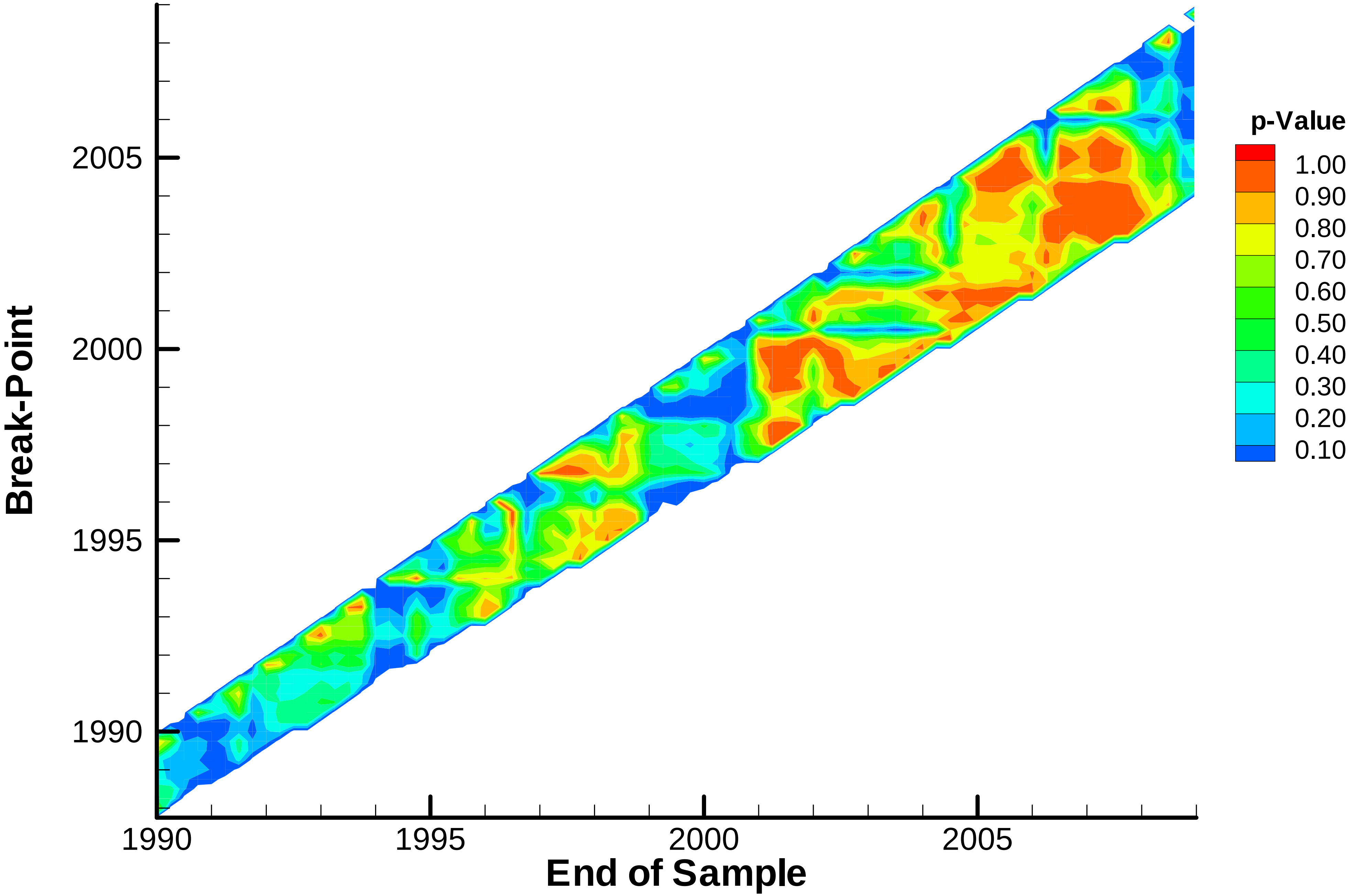


Figure 5 - End of Sample Breakpoints - (RealTime - QuasiReal)

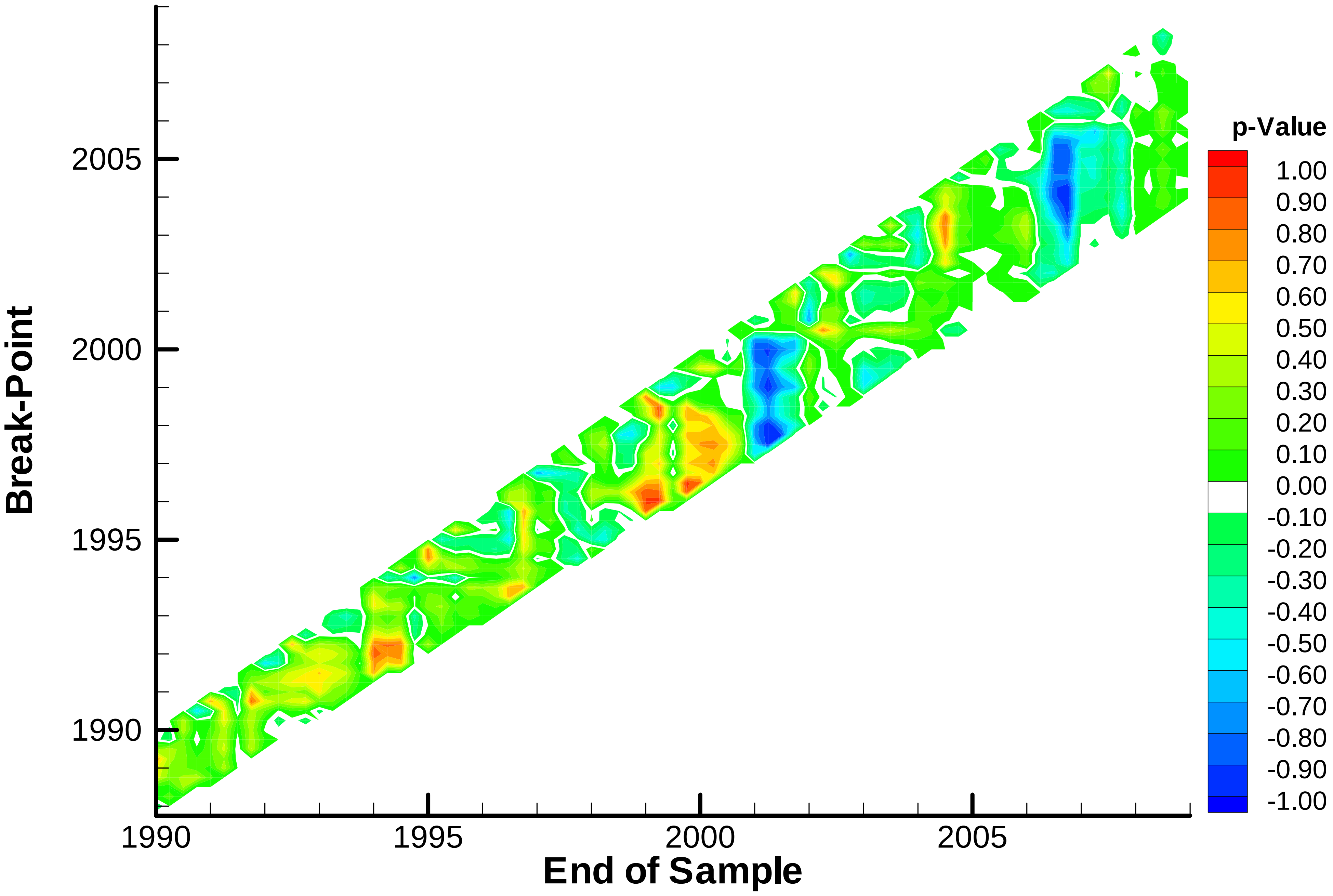
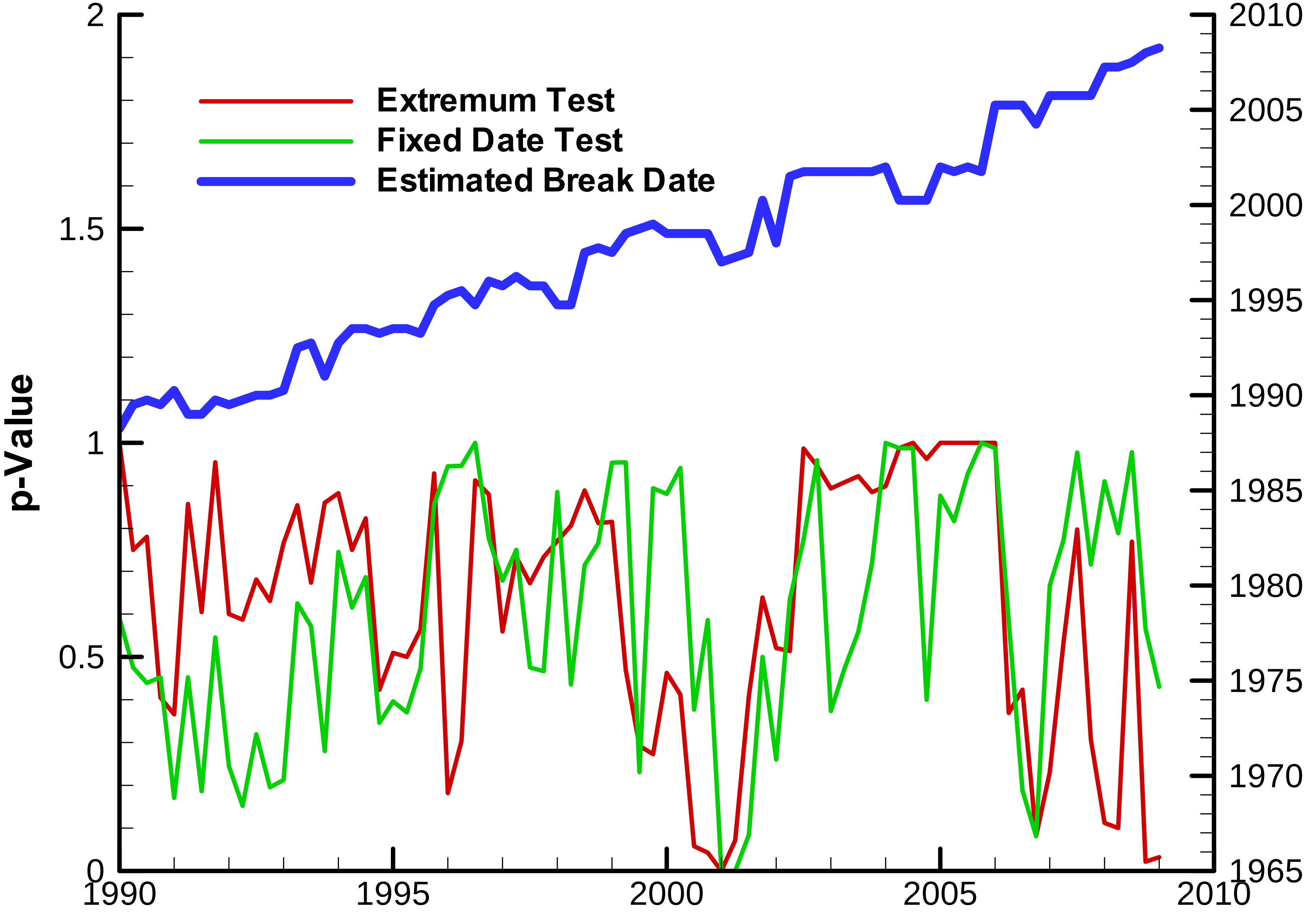


Figure 6 - Extremum and Fixed Date Tests



FigureA1: Size Distortion in Mid-Sample Tests

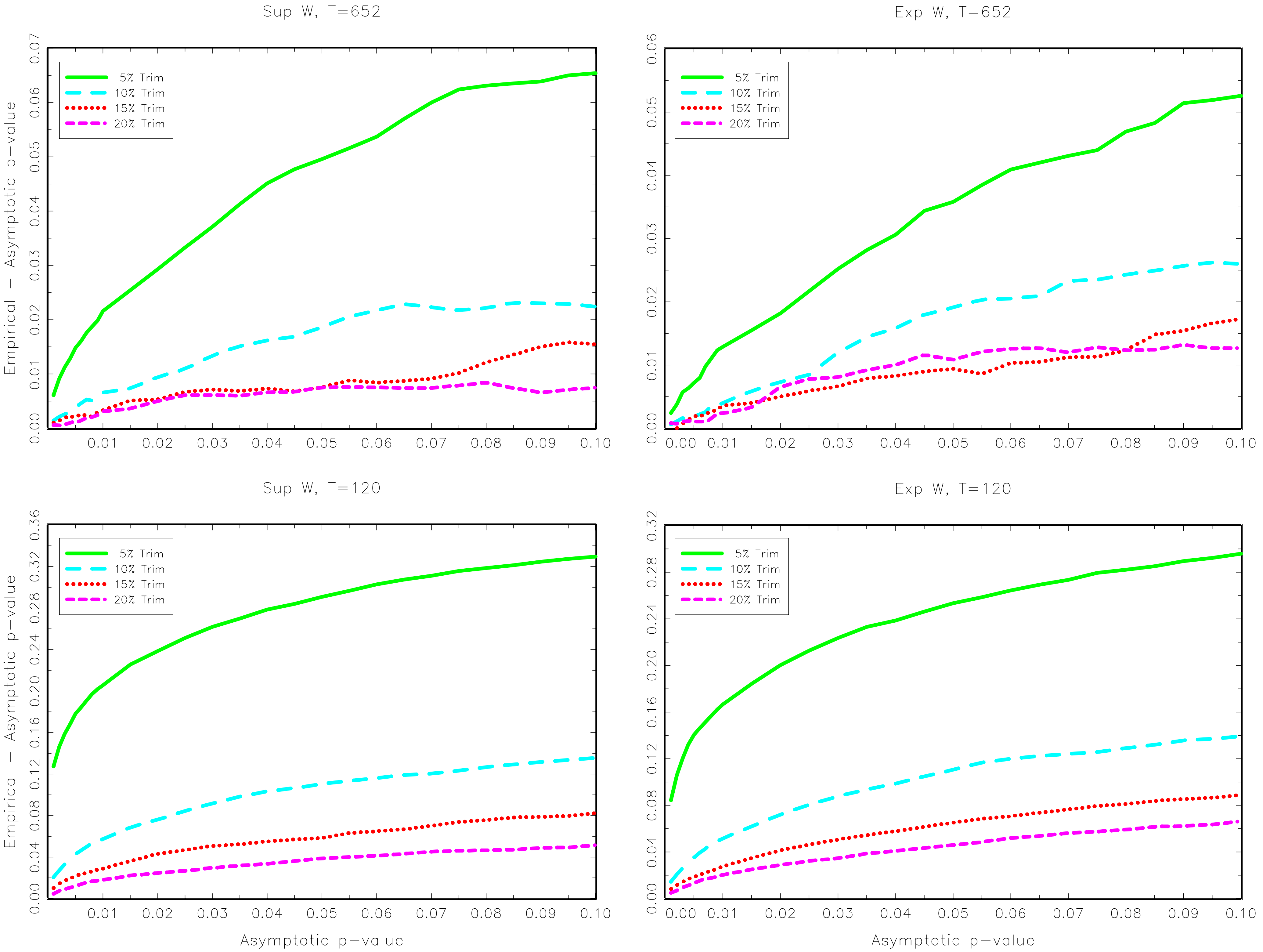


Figure A2: S-Stat Test Size

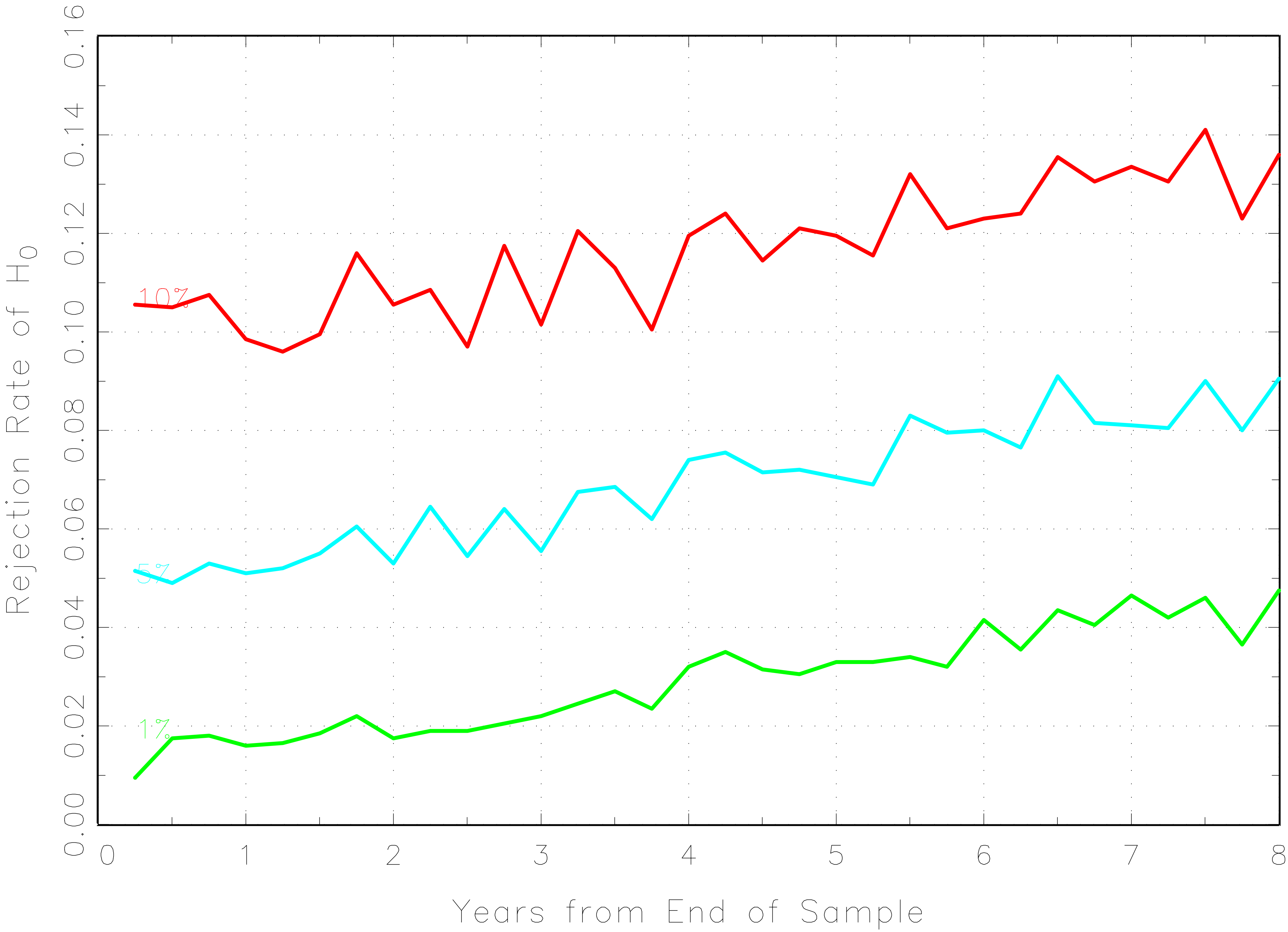
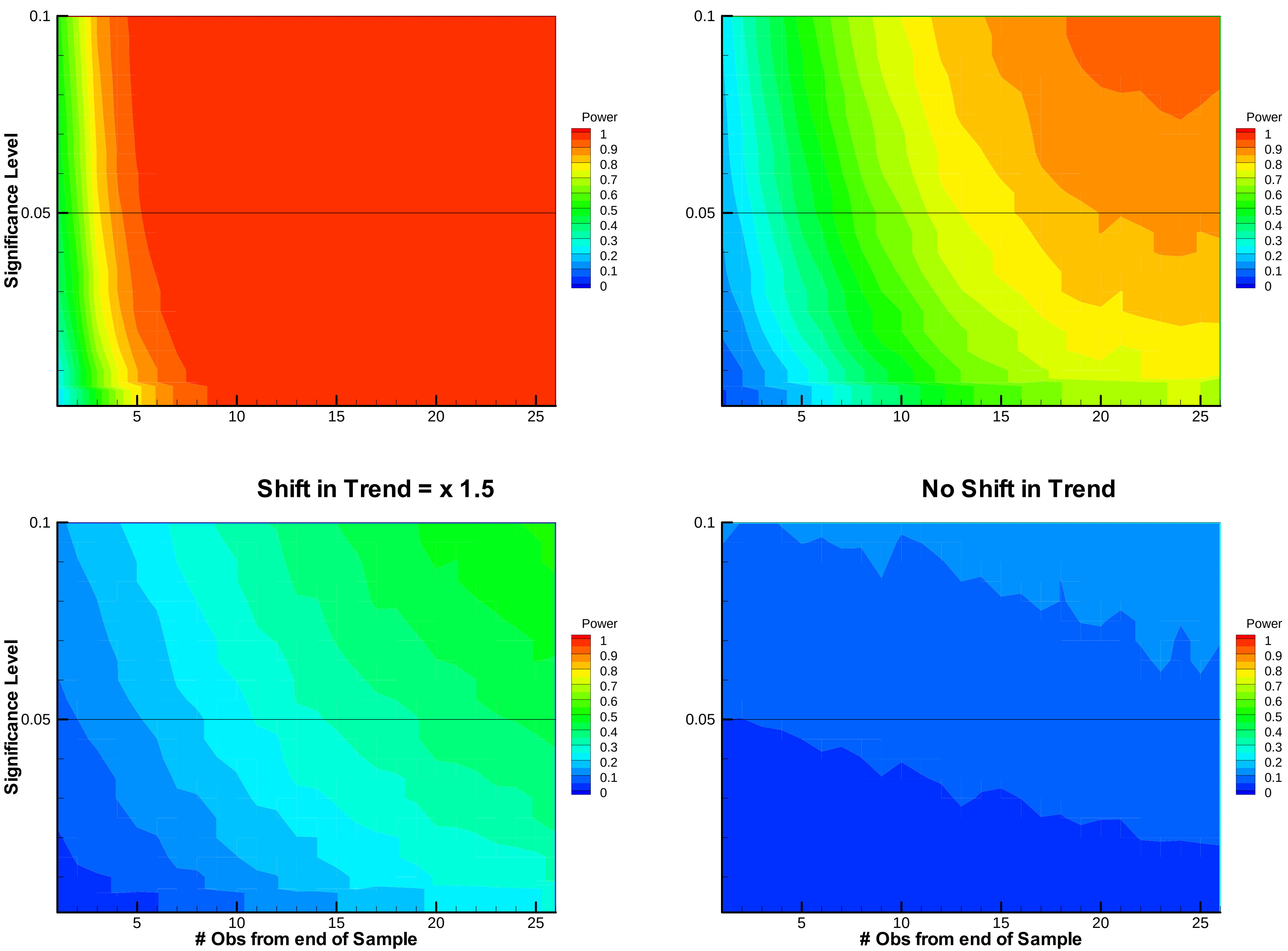


Figure A3: EOS Test Power



Shift in Trend = x 3

Shift in Trend = x 2

Shift in Trend = x 1.5

No Shift in Trend