August, 2013

Detecting structural change in university research systems: A case study of British research policy

Jian Wang
Diana Hicks, *Georgia Institute of Technology - Main Campus*

Detecting structural change in university research systems: A case study of British research policy

Jian Wang[1,2 *] & Diana Hicks[2]

[1] *Institute for Research Information and Quality Assurance (iFQ), Berlin, Germany*
[2] *School of Public Policy, Georgia Institute of Technology, Atlanta, GA, United States*

jianwang@gatech.edu
dhicks@gatech.edu

* Corresponding author

Jian Wang
685 Cherry Street
School of Public Policy
Georgia Institute of Technology
Atlanta, GA 30332-0345
United States

Phone: +1 404-884-3262
Email: jianwang@gatech.edu

Detecting structural change in university research systems: A case study of British research policy

Abstract: The university research environment has been undergoing profound change in recent decades. Aiming at international competitiveness and excellence, a variety of polices have been designed and implemented in many countries. However, evidence-based analysis of policy effects is scarce. This paper develops methods for evaluating the effect of university research policy on university system research input-output dynamics. We assume stable dynamics between inputs and outputs, and that effective policy change introduces external interventions and therefore structural changes into the system. Our proposed method involves three steps: modeling system dynamics, detecting structural change, and mapping policy change. Examining the case of the United Kingdom, we identified three structural changes in 1990, 1994, and 2001 respectively. The first change corresponds to the second round of the Research Selectivity Exercise (RSE), and the latter two changes may be caused by the database effect. The British case is a provocative demonstration of this method, which could be further developed to provide evidence-based analysis for policy learning and a foundation for cross-case comparison.

Keywords: evidence-based policy-making; performance-based research funding systems, structural change detection; research policy; UK

1. Introduction

University research is a substantial element of every national innovation system, and the university research environment has been undergoing profound change in recent decades. Aiming at higher efficiency, international competitiveness, or excellence, a variety of polices have been designed and implemented in many countries (Hicks, 2012). However, evidence-based analysis of policy effects is scarce. On the one hand, quantitative analysis is confronted with many methodological challenges; on the other hand, scientists' perceptions on potential impacts often mismatch the realities (Butler, 2010). Therefore, solid evidence-based assessments of policy effects are needed for further policy learning. The lack of solid evidence-based assessments of policy effects hampers policy learning not only in the rich countries that implemented the policies, but also in poorer countries looking for guidance on cost effective methods of improving their research systems. In response to the international need for insight into policy effectiveness, we develop methods for evaluating the effect of university research policy on university system research input-output dynamics.

The efforts of governments to boost the performance of their research systems can be understood within the canonical principal-agent model. Universities in most countries can be viewed as national systems, governed by a Ministry of Education, with professors often having civil service status. Like much of the public sector, such systems are subject to a principal-agent problem (Miller, 2005; Van der Meulen, 1998). The agents (university scientists) take actions (conduct research) that determine the payoff to the principal (the government) of its investment in research. The principal can readily observe the outcome (papers, citations, standing of universities in world rankings etc.) but not the day-to-day activities of the agents. The agent has control over the daily operations but less control over the ultimate impact of the work and is therefore inclined to report performance in terms of daily activities regardless of impact. The principal's interest is in maximizing outputs and international impact for its research investment. However, the principal faces two challenges: to screen out best agents to invest in and to monitor agent's activities.

The New Public Management (NPM) development provides partial solutions to these two problems (Herbst, 2007; Pollitt, 1993): using outputs as a mechanism to screen agents and change their incentives. The rationale for output-based funding allocation is twofold. First, funding should be given to the best performer to make the investment more efficient, and best performers can be screened out by evaluating their previous outputs. Second, linking funding to performance creates an incentive for agents to achieve better performance so that they can be more competitive in funds-seeking, which on the other hand also aligns with the principal's interest.

Based on these rationales, government core funds have been increasingly based on performance, and the funding agencies have implemented mission-oriented funding strategies and introduced output-based incentives (Heinrich & Marschke, 2010; OECD, 1998; Skoie, 1996). For example, at least fourteen countries implemented performance-based research funding systems as of 2010 (Hicks, 2012), some have introduced center of excellence competitions (China, Germany, and Japan), and others have developed national individual-level evaluations (Spain and South Africa). The universally stated goal of governments that do implement such systems has been to increase international research excellence, not in one university, but rather in their university system as a whole. However, the question remains: How effective have such systems been in overcoming the principal-agent problem and accomplishing the stated policy goal?

These policies tend to be controversial and unpopular – often accused of damaging the systems they seek to enhance. Yet rigorous assessment of these systems is in its infancy. Academics dislike the introduction of evaluation systems on principle and have therefore concentrated more on commentary than on impartial evaluation. The academic literature tends to report anecdotes or be based on surveys gathering complaints from effected scholars (Butler,

2010). However, scientists' perceptions are often found to mismatch the realities – various claimed consequences found no supporting evidence (Butler, 2010).

Therefore, evidence-based assessments on the policy effects are needed. There are some quantitative studies on how policy changes research performance. Most of them are at the national level (Butler, 2003; Furman, Murray, & Stern, 2012), and only a couple of quantitative cross-country comparison studies (Auranen & Nieminen, 2010; Franzoni, Scellato, & Stephan, 2011). While international comparative studies are needed to aid policy learning (Luwel, 2010), cross-country studies are challenging for several reasons. First, it is difficult to determine major policy change and the exact year of change to guide quantitative analysis, which requires in-depth understanding of multiple, complex, and constantly-shifting policy landscapes (Franzoni et al., 2011; Hicks, 2012). Second, it is difficult to classify countries into groups for comparison (e.g., with or without a performance-based funding system, competitive or noncompetitive system), because the systems implemented are so diverse among different countries (Geuna & Martin, 2003; Hicks, 2012; Jongbloed & Vossensteyn, 2001). Third, it is difficult to attribute observed cross-country differences in terms of research performance or efficiency to policy differences because data sources themselves suffer from cross-country variation. For example, Crespi and Geuna (2008) note that the meaning and coverage of the higher education research and development expenditure (HERD) data from the OECD are not unified for all countries, and the Thomson Reuters Web of Science (WoS) data coverage also has country biases.

Our proposed approach needs to address these challenges. Given the complexity of the timing of policy introduction, we begin simply, examining time trends in the data. This separates analysis of empirical data and of the policy context into two independent parts. Our empirical data analysis requires no knowledge about the policy context. We estimate the time of the change from the data purely based on statistical methods and then compare empirical findings with narratives of policy changes over the same time period to identify possible relationships between policies and the presence or absence of structural change. Furthermore, our empirical data analysis is conducted for each country individually, that is, empirical data are only compared between before- and after- policy change within the same country, but not across countries, and therefore the cross-country comparison challenges (i.e., the second and third challenges) are no longer relevant.

Our intention is to develop the method to study a number of countries. However, to evaluate the validity of our method, this paper analyzes only one case of the United Kingdom, which has the longest history of national research assessment and has its research policies well documented in literature. The intent of this paper is methodological. In future papers, we plan to compare policies associated with empirical structural changes and ineffective polices across all countries, in order to learn what kind of policies can be effective, and under what circumstances.

2. Approach

Time series analysis and forecasting rely on the assumption that model parameters are constant over time, while structural changes/breaks in these parameters often exist in the real-world time series data. As a result, structural change detection has been an important topic for statisticians and econometricians. Chow (1960) is the pioneer in this line of research and devised a method to use the classic *F* statistic to test whether the coefficients in two linear regressions, one before and one after the suspected break, were equal. Since then, many studies have further developed structural change detection methods: from testing a break with a known date to testing a break of unknown timing (Andrews, 1993; Andrews & Ploberger, 1994; Quandt, 1960), from testing a single break to estimating and testing multiple breaks (Bai, 1994, 1997; Bai & Perron, 1998), and from retrospectively detecting breaks in a given dataset to real-time monitoring of breaks as new data arrive (Chu, Stinchcombe, & White, 1996; Leisch, Hornik, & Kuan, 2000).

Structural change detection has proven useful in many research fields, such as analyzing the annual flow of the Nile river (Cobb, 1978), finding evidence of global warming from temperature time series (Fomby & Vogelsang, 2002; Vogelsang & Franses, 2005), testing impacts of the 1929 Great Crash and the 1973 oil price shock on the US economy (Perron, 1989), analyzing "slowdowns" and "meltdowns" of national economies (i.e., GDP) (Ben-David & Papell, 1998), identifying breaks in US labor productivity (Hansen, 2001), and testing some economic theories predicting changes in trade volumes over time (Bunzel & Vogelsang, 2005).

However, structural change detection has not been used in the bibliometrics community. The OECD started collecting research and development (R&D) data for its member countries on a regular basis since the 1960s (OECD, 2010), and the Institute for Scientific Information (ISI, now part of Thomson Reuters) was begun in 1955 by Dr. Eugene Garfield to systematically index scientific publications (Thomson Reuters, 2013). As ever longer time series of science and technology data accumulate, structural change detection becomes more obviously useful as a powerful tool to analyze dynamics and breaks in research systems. This paper is a first attempt to use this method to test policy effects on national research systems.

We assume that (1) a national university research system has stable dynamics between research inputs and outputs and (2) an effective policy is an external intervention which introduces structural change into the system. Different from classical policy evaluation approaches which focus on a specific policy and aim to assess impacts of this focal policy, our approach starts with evidence-based empirical data analysis to determine whether or not a country's science system exhibits a structural change in the relationship between input and output over the past three decades. We then compare the empirical results with narratives of the development of university science policy over the same time period to identify possible relationships between policies and the presence or absence of structural change. This approach consists of three components: (1) modeling system dynamics, (2) detecting structural change,

and (3) mapping policy change. We take the UK for a case study to test the reliability of our approach, because policy changes in the UK university system are well-documented.

We analyze three decades of publication and funding data. Research output is measured by the number of Thomson Reuters Web of Science (WoS) indexed journal publications (articles, letters, notes, and reviews), with at least one British university affiliation. We have publication counts from 1981 to 2011. Research input is measured by the annual higher education research and development expenditure (HERD) in constant 2005 dollars and discounted for purchasing parity power (PPP) collected from the OECD. UK HERD data from 1981 to 2010 are available.

3. Modeling system dynamics

To justify public investments in basic research, there has been a longstanding interest in measuring general economic benefits from public science investments (Lane & Bertuzzi, 2011; Mansfield, 1991; Narin, Hamilton, & Olivastro, 1997; Stephan, 1996), and several studies have modeled the input-output dynamics of national research systems (Adams & Griliches, 1998; Crespi & Geuna, 2008; Johnes & Johnes, 1995). Following Griliches (1979) and Crespi and Geuna (2008), we adopt a knowledge production function, $Y = F(X, u)$, to describe the relationship between research output (Y) and input (X and u), where X represents current and past R&D expenditures, and u represents all other unmeasured factors. Furthermore, we adopt the commonly used Cobb–Douglas functional form and assume that u is random after adding the time variable to measure systematic components of the unmeasured factors, that is,

$$Y_t = A \cdot X_t^{\beta} \cdot e^{\lambda \cdot t + u_t} \quad (1)$$

where t is a time index, A is a constant, e is the natural logarithm base, and β and λ are parameters to be estimated. Furthermore, to take into account of the effect of past R&D expenditures on current research output, it is assumed that

$$X_t = \prod_{i=0}^{q} R_{t-i}^{w_i} \quad (2)$$

where R_{t-i} is the R&D expenditure in year t-i, q is the maximum lag (i.e., R&D expenditures before year t-q are assumed to have no effect on current knowledge production), and w is the weight. Both q and w are fixed parameters to be estimated. Then we plug (2) into (1) and take natural logarithms on both sides and get

$$P_t = \alpha + \beta \cdot \sum_{i=0}^{q} w_i \cdot H_{t-i} + \lambda \cdot t + u_t \quad (3)$$

where P is the natural log of current research output (number of publications), α is the log of A,

and H is the log of HERD.

To select the optimal q and estimate α, β, w, and λ, we follow the same procedure as Crespi and Geuna (2008). q is selected using the Akaike Information Criteria (AIC), the Bayesian information criterion (BIC), and the Bozdogan Index of Information Complexity (ICOMP). For a more parsimonious model, an Almon model is subsequently fitted (Almon, 1965), which assumes that the lag weights, w_i, follow a polynomial going back q years and then become zero, that is, imposes the following constraints:

$$w_i = \sum_{j=0}^{m} c_j \cdot i^j \tag{4}$$

where m is the degree of polynomial, which can be selected following a backward elimination procedure and F-tests, and c_j is estimated from linear regression.

Following this procedure, we find the optimal q to be 6 and m to be 1. Further details are reported in Appendix 1. In addition, we fit another model with the optimal parameters identified in Crespi and Geuna (2008), that is, $q = 6$ and $m = 3$. We also fit an unrestricted polynomial distributed lag (PDL) model, that is, the model described by equation (3) without imposing constraints of (4). Finally, we fit a model without HERD. Therefore, we have the following four models, and model fitting results are reported in *Table 1*.

$$Mod\ 1: P_t = \alpha + \lambda \cdot t + u_t \tag{5}$$

$$Mod\ 2: P_t = \alpha + \beta \cdot \sum_{i=0}^{6} w_i \cdot H_{t-i} + \lambda \cdot t + u_t \tag{6}$$

$$Mod\ 3: P_t = \alpha + \beta \cdot \sum_{i=0}^{6} w_i \cdot H_{t-i} + \lambda \cdot t + u_t,\ where\ w_i = \sum_{j=0}^{3} c_j \cdot i^j \tag{7}$$

$$Mod\ 4: P_t = \alpha + \beta \cdot \sum_{i=0}^{6} w_i \cdot H_{t-i} + \lambda \cdot t + u_t,\ where\ w_i = \sum_{j=0}^{1} c_j \cdot i^j \tag{8}$$

4. Detecting structural change

The next step is to detect structural changes in the four linear regression models described above. A structural change in year t would mean that the regression coefficients (i.e., α, β, w, and λ) would change, and therefore that we should partition the time series into two segments and fit two different models for data before and after t separately. Structural change detection involves two steps: (1) estimate the year of breaks assuming there are n breaks, and (2) test how many breaks are significant, in other words, choose n.

In the first step, estimating the year of given n (n = 1, 2, …) structural changes, we evaluate all possible combinations of n breakpoints, partitioning the time series into $n + 1$ segments, and for each n select the partioning that minimizes the sum of squared residuals. In the second step to choose n, we adopt the sequential testing procedure proposed by Kejriwal and Perron (2010). We sequentially examine the $F(l+1|l)$ statistics and select n such that the tests $F(l+1|l)$ are insignificant for $l \geq n$.

Table 1. Four model fitting results

	Mod 1	Mod 2	Mod 3	Mod 4
H_t		-0.693[0.271]**	-0.436[0.192]**	-0.333[0.081]***
H_{t-1}		0.009[0.414]	-0.125[0.102]	-0.251[0.060]***
H_{t-2}		0.086[0.507]	-0.089[0.112]	-0.168[0.042]***
H_{t-3}		-0.717[0.537]	-0.166[0.069]**	-0.085[0.035]**
H_{t-4}		0.410[0.505]	-0.198[0.120]	-0.003[0.045]
H_{t-5}		-0.422[0.437]	-0.024[0.108]	0.080[0.063]
H_{t-6}		0.637[0.299]*	0.515[0.208]**	0.162[0.085]*
Year	0.040 [0.001]***	0.073[0.011]***	0.065[0.010]***	0.067[0.010]***
Intercept	-68.640 [2.041]***	-128.600[20.260]***	-113.894[18.545]***	-118.238[18.358]***
R2 adj	0.981	0.979	0.980	0.979
Obs. #	31	24	24	24
Constraints (χ^2)			0.601	4.800

Standard errors reported in brackets. (*) significant at 10%; (**) significant at 5%; (***) significant at 1%. H_t is the natural log of HERD in year t. The dependent variable is the natural log of publication counts in year t. Mod1, 2, 3, and 4 are described by equation (5), (6), (7), and (8) respectively. Mod 1 is the model without HERD; Mod 2 is an unrestricted polynomial distributed lag (PDL) model with a lag length of six. Mod 3 is an Almon model with a lag length of six and a polynomial degree of three, and Mod 4 is an Almon model with a lag length of six and a polynomial degree of one.

Table 2. Structural change detection in four hypothetical cases

	S1	S2	S3	S4
True Model				
Intercept 1	-40.000	-40.000	-90.000	-90.000
Intercept 2	-139.800	-40.100	-140.000	-60.160
t 1	0.025	0.025	0.050	0.050
t 2	0.075	0.025	0.075	0.035
Without Breaks				
Intercept	-87.417[4.840]***	-25.290[1.791]***	-103.900[2.817]***	-66.099[1.802]***
T	0.049[0.002]***	0.018[0.001]***	0.057[0.001]***	0.038[0.001]***
R2 adj	0.933	0.930	0.983	0.984
With Breaks				
Intercept 1	-39.780[1.779]***	-39.780[1.779]***	-89.780[1.779]***	-89.780[1.779]***
Intercept 2	-141.400[1.792]***	-41.730[1.792]***	-141.600[1.792]***	-61.740[1.792]***
t 1	0.025[0.001]***	0.025[0.001]***	0.050[0.001]***	0.050[0.001]***
t 2	0.076[0.001]***	0.026[0.001]***	0.076[0.001]***	0.036[0.001]***
R2 adj	1.000	1.000	1.000	1.000
1 break	1995	1995	1995	1995
2 breaks	1996, 2004	1995, 2000	1995, 2000	1995, 2000
3 breaks	1996, 1999, 2004	1995, 1999, 2004	1995, 1999, 2004	1995, 1999, 2004
F(1|0)	∞***	7.257***	20.262***	7.765***
F(2|1)	0.309	0.309	0.309	0.309
F(3|2)	2.132	1.146	1.146	1.146

Standard errors reported in brackets. (*) significant at 10%; (**) significant at 5%; (***) significant at 1%. "True Model" specifies the true parameters. In all four cases, the random error $\varepsilon_t \sim_{iid} N(0, 0.015^2)$. The "Without Breaks" section reports least square regression estimated coefficients assuming no structural change in the system, and the "With Breaks" section reports regression results after incorporating the detected structural change. The row of "1 break" reports the estimated year of change when assuming there is 1 break in the system. F(3|2) reports the test statistics testing the null hypothesis of 2 changes versus the alternative hypothesis of 3 changes.

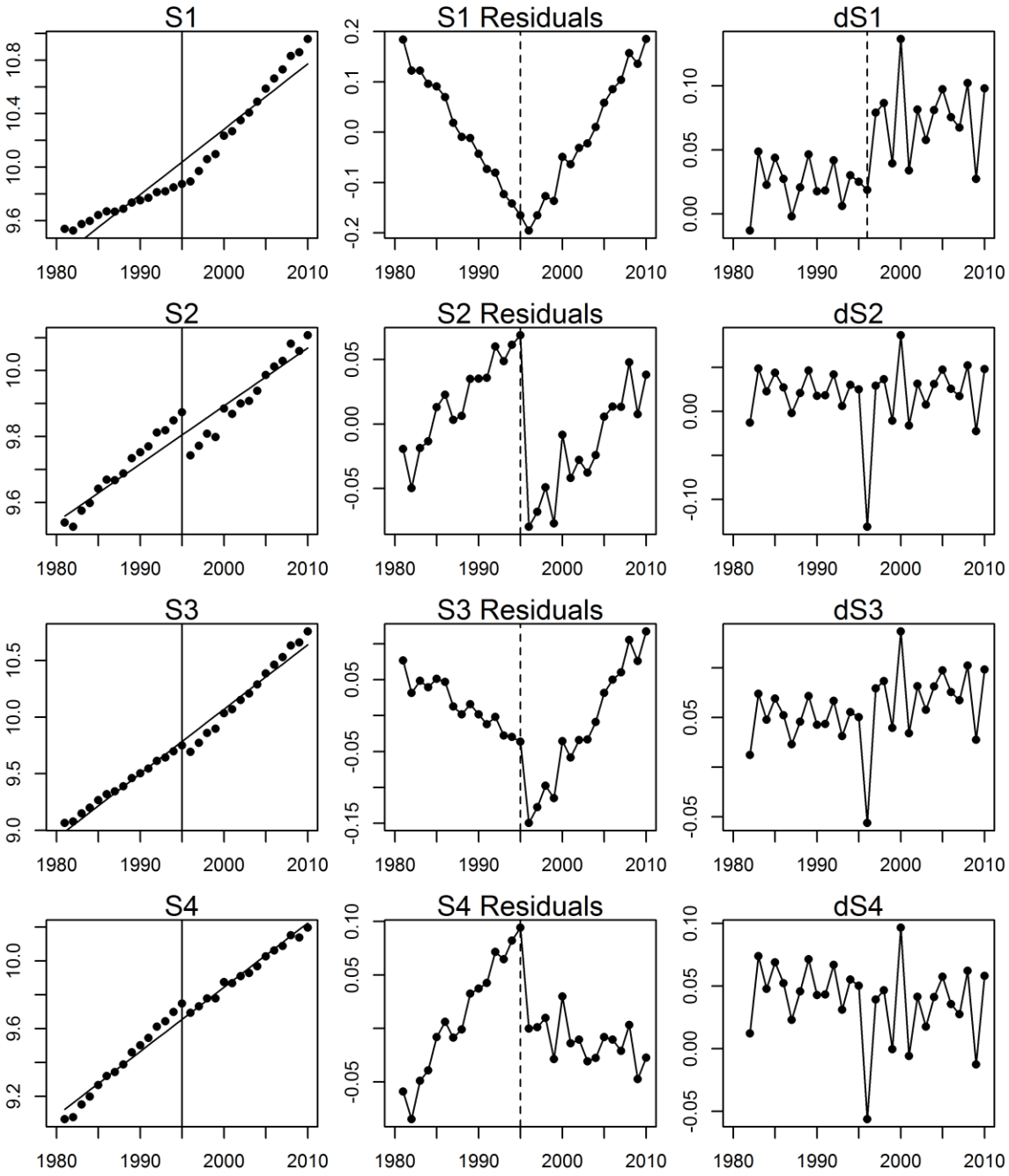

Figure 1. Four hypothetical cases. The horizontal axes are year *t*. Points in "S1" are simulated time series *S1*, the vertical solid line is the true breakpoint. The upward sloping straight line is the regression line assuming no structural change. The differences between the true values and the fitted regression line are residuals. These residuals are plotted in "S1 Residuals." "dS1" plots first-differenced data, that is, $dS1_t = S1_t - S1_{t-1}$. The vertical dashed lines are detected structural changes. Analyzing *dS2, dS3,* and *dS4* would fail to detect the true change.

In addition, residuals of the regression models assuming no breaks are plotted to aid visual evaluation. As shown in *Table 2*, four hypothetical time series from 1981 to 2011 are simulated, each with a structural change between 1995 and 1996 (i.e., the time series are truly partitioned into two segments: [1981, 1995] and [1996, 2011]). Note that to comply with the convention in literature, we notate the time of break as *t* when there is a break between *t* and *t* + 1, that is, when we say the year of break is *t*, the breakpoint *t* belongs to the segment before the change. Taking *S1* as an example, the true model is:

$$S1_t = \begin{cases} -40.000 + 0.025 \cdot t + \varepsilon_t, & t \in [1981, 1995] \\ -139.800 + 0.075 \cdot t + \varepsilon_t, & t \in [1996, 2011] \end{cases} \quad (9)$$

where *t* is a year index, and the random errors ε_t are independently and identically distributed following a normal distribution with mean 0 and standard deviation 0.015. *S1* exhibits a change in the slope of the trend without a change in the level of the trend, that is, the trend function is joined at the time of break. If we ignore this structural change and fit a linear regression model for the whole time period, then we would force two segments into one straight line (i.e., the regression line in the "S1" plot in *Figure 1*), and the residuals (i.e., true values minus fitted values) will first increase and then decrease, as shown in the "S1 Residuals" plot in *Figure1*. Three other generic scenarios are: S_t with a change in the intercept but not the slope of the trend (*S2*), S_t with both changes in the slope (increase) and the level (decrease) of the trend (*S3*), and S_t with both changes in the slope (decrease) and the level (decrease) of the trend (*S4*). In summary, we will observe a change in the direction (e.g., from increase to decrease) of the residuals over time when there is a slope change in the trend (*S1*, *S3*, and *S4*), and a dramatic jump in the residuals when an intercept change is present (*S2*, *S3*, and *S4*). These patterns will help visually evaluate structural changes.

After the visual evaluation, we conduct the two steps of structural change detection: estimate breakpoints assuming there are 1, 2, and 3 breaks respectively, and then use the $F(l+1|l)$ statistics to select the number of breaks through sequentially testing the null hypothesis of l changes versus the alternative hypothesis of l+1 changes. Structural change detection results are reported in *Table 2*. Take *S1* as an example, the breakpoint is 1995 if we assume one change, 1996 and 2004 if we assume two changes. The tests suggest that only one change is significant. Therefore, the conclusion is that there is one change in 1995.

In addition, a common practice is to analyze the first-differenced data or growth rates, which removes the trend from the data. In other words, $dS_t = S_t - S_{t-1}$, instead of S_t itself, is analyzed. However, Perron and colleagues have argued that aggregate macroeconomic time series are better modeled as being stationary around a broken trend (that is, the trend function has some breaks in its level and/or slope), so performing a structural change test using first-

differenced data is not appropriate (Kejriwal & Perron, 2010; Perron, 1989; Perron & Yabu, 2009a, 2009b). We also found that using first-differenced data will fail to detect changes when the variance of the random errors is relatively large (*S3* and *S4*) or only the intercept is changed (*S2*). In addition, both the sequential tests developed by Kejriwal and Perron (2010) for testing breaks in the trend function and the sequential tests developed by Bai and Perron (1998) for trend-stationary data were adopted for analyzing first-differenced data, and they gave the same conclusions. Therefore, this first-difference approach is not adopted.

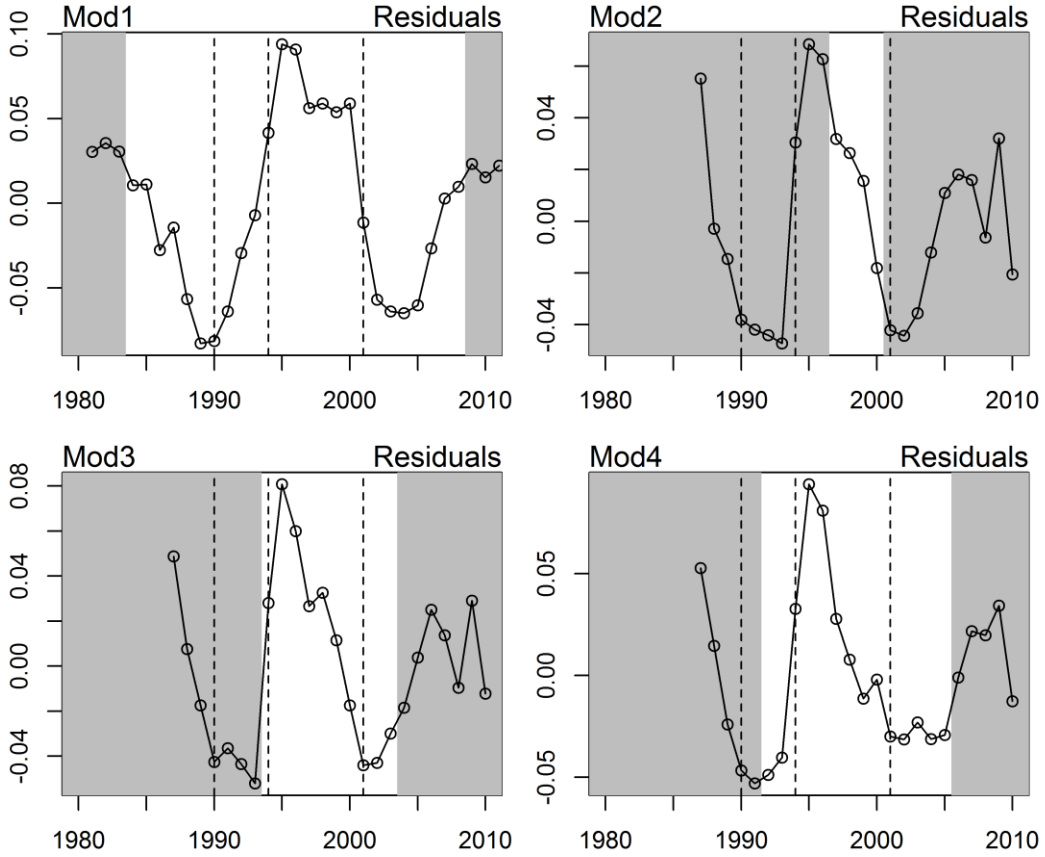

Figure 2. Residuals of four models without break. The horizontal axes are year. Grey areas are not testable. The vertical dashed lines are detected structural changes in Mod 1.

Table 3. Mod 1 structural changes estimations and tests

l	Breaks	*F*(*l*|*l*-1)
1	1992	0.172
2	1991, 2000	3.691*
3	1990, 1994, 2001	8.722***
4	1990, 1994, 1999, 2002	3.936
5	1990, 1994, 1999, 2002, 2006	2.263

F(k|k-1) test the null hypothesis of *k-1* breaks versus the alternative hypothesis of *k* breaks. (*) significant at 10%; (**) significant at 5%; (***) significant at 1%.

Structural change detection results on the British data are reported in *Table 3*. Three significant changes are detected in Mod1: 1990, 1994, and 2001. The power to detect changes in Mod 2, 3, and 4 is however very limited, for two major reasons. First, they incorporate six-year lagged terms of *ln*(HERD), so we lose six data points and have data from 1987 to 2011 available for analysis. Second, these models have more predictors in the model, so the minimum length of a stable time period (i.e., assuming without a change) required by linear regression is longer. For example, we have nine parameters to be estimated in Mod 2, so each segment requires at least ten data points to be able to fit a regression model for that segment. Therefore, the only testable period for Mod 2 is between 1996 and 2000 (i.e., the white area between two grey sections in the "Mod 2 Residuals" plot in *Figure 2*).

Table 4. Mod 1 with three detected structural changes

	Model with breaks
Intercept [1981, 1990]	-39.980 [3.414]***
Intercept [1991, 1994]	-136.100 [13.920]***
Intercept [1995, 2001]	-41.320 [5.898]***
Intercept [2002, 2011]	-92.350 [3.450]***
Year [1981, 1990]	0.025 [0.002]***
Year [1991, 1994]	0.074 [0.007]***
Year [1995, 2001]	0.026 [0.003]***
Year [2002, 2011]	0.052 [0.002]***
R2 adj	1
Obs. #	31

Standard errors reported in brackets. (*) significant at 10%; (**) significant at 5%; (***) significant at 1%.

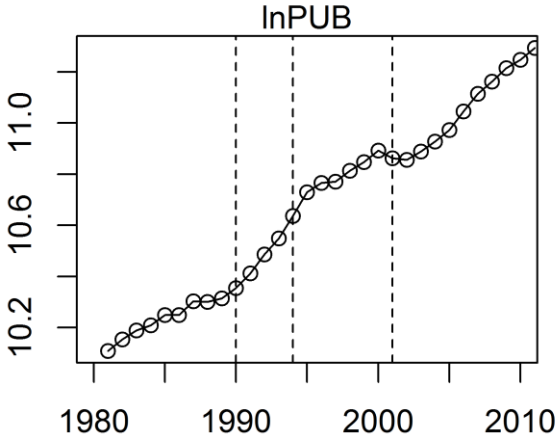

Figure 3. Natural log of publication output by year. The vertical dashed lines are detected structural changes in Mod 1.

Given this limitation, our interpretation strategy is to mainly use results of Mod 1 and then use visualization to assess whether the changes detected in Mod 1 vanish as we add HERD and its lagged terms to the model. *Table 4* and *Figure 3* report three major changes detected in Mod 1. The number of publications grew much faster after 1991 and then slowed down again in 1995. It switched back to faster growth in the period since 2002. These changes are largely untestable in Mod 2, 3, and 4, but they seem still apparent (*Figure 2*), that is, the residuals still exhibit a direction-changing pattern at these points. Therefore, we conclude three structural changes in the British university research system.

5. Mapping policy change

The change in 1990 coincides with the second round of the Research Selectivity Exercise (RSE, as the Research Assessment Exercise was initially called). The first university research assessment in the UK was conducted in 1986, the second in 1989, the third in 1992. That the break seems related to the second and not first round of the evaluation is interesting. It makes quite a bit of sense, given changes between the first and second rounds. Martin and Whitley (2010, pp. 54-57) discuss the evolution of the RSE and the differences between the 1986 and 1989 versions. In the 1986 round, departments submitted details of their five best publications from the previous five years. The effect on funding was rather limited and "some in the more established universities paid relatively little attention (hoping, no doubt, that the RSE would 'go away'), others took it much more seriously" (Martin & Whitley, 2010, p. 55). This suggests a limited effect on university behavior, which is consonant with the lack of a break in the publication trend data. In 1989, departments submitted details on up to two publications per faculty member (raised to 4 in 1996) as well as the total number of publications in relation to full-time staff. Further, the results of the evaluation were more explicitly linked to a larger amount of funding – half of the research portion of the block grant was allocated on the basis of the 1989 ratings (Martin & Whitley, 2010, pp. 56-57). The increased importance of each individual's productivity in the ranking, as well as the greater financial stakes, and no doubt the sense that this was not "going away," all suggest a more substantial impact on faculty behavior, aligned with the shift in UK university publications to a faster growth trajectory beginning in 1989. In 1992 departments were allowed to submit information only for research active staff. Thereafter, tweaks were made, but the method had settled down.

The breakpoint analysis suggests not only that the RAE policy was strong enough to shift the university system in the desired direction, but that the design of the system mattered and that the big shift was seen when individual level productivity began to matter and when significant money began to move. Unfortunately, it is not possible to disentangle the effects of these two factors in this instance. The conclusion that individual level attention is needed to create a shift is aligned with the findings of Franzoni et al. (2011).

The slowing in 1995 and acceleration in 2001 may be explained by a database effect. One of the challenges in looking for changes in scientific output due to policy implementation is that scientific output is measured in a database, and Thomson Reuters can choose to enlarge the database or not for commercial reasons. So the database is subject to change for reasons entirely unconnected with national policy shifts. To control for this effect, we used the same techniques to find breakpoints in the growth of number of papers indexed in WoS. We found that database growth slowed around 1995 and accelerated around 2001, and these breaks appear in the publication trends of the core English language countries whose publications are favored in WoS coverage.

6. Discussion

We propose this method in response to the international need for evidence-based assessments of policy effects, and the British case study demonstrate the reliability of this method. This method could be further developed to provide evidence-based analysis for policy learning and a foundation for cross-case comparison. For future studies, we could apply this method to other countries, detect structural changes in a much larger sample of countries, and then identify effective polices that contributed to these detected changes. Such study could provide rich information for policy learning, and inform more effective and efficient policy practice in both rich and poor countries.

However, this method has several limitations. First, the method is largely constrained by the quantity and quality of the data available. OECD HERD data are available after 1981 for some countries and not available for many other countries, and not all countries' data are robust (Crespi & Geuna, 2004). Adding lagged HERD to the model seriously shortens the length of time series and increases the length of untestable periods. Other problems beset trends in WoS data. As discussed before, the WoS database is subject to change for reasons entirely unconnected with national policy shifts. Therefore, data limits may constrain the application of structural change detection, and the problem of database change may introduce biases in modeling system dynamics and detecting structural changes.

Second, this method assumes a stable knowledge production system or a smooth exponential growth in national publication counts, with limited number of breaks. However, if there are constant changes in very short periods, then our assumption is violated and the detecting method is problematic. In addition, if the growth of publication counts is not exponential, then a different model specification for Mod 1 is needed. Third, concluding a causal relationship between a possible policy and the identified structural change might present another challenge and require significant inputs from national experts.

The first limitation seems to be the most important and relevant, so we propose the following procedure to address the data shortage problem. We assume that HERD is growing

smoothly and the effect of HERD is also stable. Based on this assumption we use Mod 1 to analyze the growth of publication counts without accounting for HERD, and detect structural changes in the system. Subsequently, we add HERD to the model and use residual visualization to assess whether these detected changes are still apparent. If they vanish, then they can be explained by the change in HERD. If they are still apparent, then these changes are not due to funding change, and therefore we can conclude changes in system dynamics.

To address the database change problem, we propose the following two possible strategies. One strategy is to conduct the structural change detection analysis on the whole database before individual-country analysis and cross-country comparison, and breaks identified in countries can be attributed to database effect (not policy effect) through comparing country analysis results and database analysis results. This strategy is implemented in this paper. Another strategy is to integrate database change into modeling system dynamics by adding variables to the linear regression models to control for database effects.

In addition, this methodological paper analyzed only the case of the United Kingdom to evaluate the validity of our method. For future studies, we could apply this method to a much larger sample of countries for cross-country comparisons. In addition to this extension regarding to sample of countries, we should also investigate many other aspects of research performance. For example, we can look into citation impacts and detect subtle change in citation distributions, to assess if national policies have successfully improved research excellence, and we can also examine portfolios of national research to investigate whether these policies have unintended consequences of shifting research to fields with higher publication or citation rates in order to gain favorable evaluation outcomes.

Acknowledgment

The authors would like to thank Stefan Hornbostel, Sybille Hinze, Linda Butler, Rainer Frietsch, and two anonymous reviewers for their helpful comments and suggestions. The data used in this paper are from a bibliometrics database developed and maintained by the Competence Centre for Bibliometrics for the German Science System (KB) and derived from the 1980 to 2011 Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA: © Copyright Thomson Reuters (Scientific) 2012. The authors thank the KB team for its collective effort in the development of the KB database.

References:

Adams, James D., & Griliches, Zvi. (1998). Research Productivity in a System of Universities. *Annals of Economics and Statistics / Annales d'Économie et de Statistique*(49/50), 127-162.

Almon, Shirley. (1965). The Distributed Lag Between Capital Appropriations and Expenditures. *Econometrica, 33*(1), 178-196. doi: 10.2307/1911894

Andrews, D. W. K. (1993). TESTS FOR PARAMETER INSTABILITY AND STRUCTURAL-CHANGE WITH UNKNOWN CHANGE-POINT. *Econometrica, 61*(4), 821-856. doi: 10.2307/2951764

Andrews, D. W. K., & Ploberger, W. (1994). OPTIMAL TESTS WHEN A NUISANCE PARAMETER IS PRESENT ONLY UNDER THE ALTERNATIVE. *Econometrica, 62*(6), 1383-1414. doi: 10.2307/2951753

Auranen, Otto, & Nieminen, Mika. (2010). University research funding and publication performance—An international comparison. *Research Policy, 39*(6), 822-834. doi: 10.1016/j.respol.2010.03.003

Bai, Jushan. (1994). LEAST SQUARES ESTIMATION OF A SHIFT IN LINEAR PROCESSES. *Journal of Time Series Analysis, 15*(5), 453-472. doi: 10.1111/j.1467-9892.1994.tb00204.x

Bai, Jushan. (1997). Estimating multiple breaks one at a time. *Econometric Theory, 13*(3), 315-352.

Bai, Jushan, & Perron, Pierre. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica, 66*(1), 47-78. doi: 10.2307/2998540

Ben-David, D., & Papell, D. H. (1998). Slowdowns and meltdowns: Postwar growth evidence from 74 countries. *Review of Economics and Statistics, 80*(4), 561-571. doi: 10.1162/003465398557834

Bunzel, H., & Vogelsang, T. I. (2005). Powerful trend function tests that are robust to strong serial correlation, with an application to the prebisch-singer hypothesis. *Journal of Business & Economic Statistics, 23*(4), 381-394. doi: 10.1198/073500104000000631

Butler, Linda. (2003). Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. *Research Policy, 32*(1), 143-155. doi: 10.1016/s0048-7333(02)00007-0

Butler, Linda. (2010). Impacts of performance-based research funding systems: a review of the concerns and the evidence. In OECD (Ed.), *Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings* (pp. 127-165). Paris: OECD Publishing.

Chow, G.C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society, 28*(3), 591-605.

Chu, C. S. J., Stinchcombe, M., & White, H. (1996). Monitoring structural change. *Econometrica, 64*(5), 1045-1065. doi: 10.2307/2171955

Cobb, George W. (1978). The Problem of the Nile: Conditional Solution to a Changepoint Problem. *Biometrika, 65*(2), 243-251. doi: 10.2307/2335202

Crespi, Gustavo A., & Geuna, Aldo. (2004). *The Productivity of Science*. Brighton: SPRU, University of Sussex.

Crespi, Gustavo A., & Geuna, Aldo. (2008). An empirical study of scientific production: A cross country analysis, 1981–2002. *Research Policy, 37*(4), 565-579. doi: 10.1016/j.respol.2007.12.007

Fomby, T. B., & Vogelsang, T. J. (2002). The application of size-robust trend statistics to global-warming temperature series. *Journal of Climate, 15*(1), 117-123. doi: 10.1175/1520-0442(2002)015<0117:taosrt>2.0.co;2

Franzoni, Chiara, Scellato, Giuseppe, & Stephan, Paula. (2011). Changing incentives to publish science. *Science, 333*(6043), 702-703.

Furman, Jeffrey L., Murray, Fiona, & Stern, Scott. (2012). Growing Stem Cells: The Impact of Federal Funding Policy on the U.S. Scientific Frontier. *Journal of Policy Analysis and Management, 31*(3), 661-705. doi: 10.1002/pam.21644

Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva, 41*(4), 277-304.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics, 10*(1), 92-116.

Hansen, B. E. (2001). The new econometrics of structural change: Dating breaks in US labor productivity. *Journal of Economic Perspectives, 15*(4), 117-128. doi: 10.1257/jep.15.4.117

Heinrich, Carolyn J., & Marschke, Gerald. (2010). Incentives and their dynamics in public sector performance management systems. *Journal of Policy Analysis and Management, 29*(1), 183-208. doi: 10.1002/pam.20484

Herbst, M. (2007). *Financing public universities: The case of performance funding* (Vol. 18). Dordrecht: Springer.

Hicks, Diana. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251-261. doi: 10.1016/j.respol.2011.09.007

Johnes, Jill, & Johnes, Geraint. (1995). Research funding and performance in U.K. University Departments of Economics: A frontier analysis. *Economics of Education Review, 14*(3), 301-314. doi: 10.1016/0272-7757(95)00008-8

Jongbloed, Ben, & Vossensteyn, Hans. (2001). Keeping up Performances: An international survey of performance-based funding in higher education. *Journal of Higher Education Policy and Management, 23*(2), 127-145. doi: 10.1080/13600800120088625

Kejriwal, M., & Perron, P. (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*(5), 305-328. doi: 10.1111/j.1467-9892.2010.00666.x

Lane, Julia, & Bertuzzi, Stefano. (2011). Measuring the results of science investments. *Science, 331*(6018), 678-680. doi: 10.1126/science.1201865

Leisch, F., Hornik, K., & Kuan, C. M. (2000). Monitoring structural changes with the generalized fluctuation test. *Econometric Theory, 16*(6), 835-854. doi: 10.1017/s0266466600166022

Luwel, Marc. (2010). Highlights and reflections: Rapporteur's report. In OECD (Ed.), *Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings* (pp. 167-174). Paris: OECD Publishing.

Mansfield, Edwin. (1991). Academic research and industrial innovation. *Research Policy, 20*(1), 1-12. doi: 10.1016/0048-7333(91)90080-a

Martin, B.R., & Whitley, R. (2010). The UK Research Assessment Exercise: A Case of Regulatory Capture? In R. Whitley, J. Glaser & L. Engwall (Eds.), *Reconfiguring Knowledge Production: Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation* (pp. 51-80). Oxford: Oxford University Press.

Miller, Gary J. (2005). The political evolution of principal-agent models. *Annual Review of Political Science, 8*(1), 203-225. doi: 10.1146/annurev.polisci.8.082103.104840

Narin, Francis, Hamilton, Kimberly S., & Olivastro, Dominic. (1997). The increasing linkage between U.S. technology and public science. *Research Policy, 26*(3), 317-330. doi: 10.1016/s0048-7333(97)00013-9

OECD. (1998). *University Research in Transition*. Paris: OECD.

OECD. (2010). Main Science and Technology Indicators. Retrieved May 21, 2013, from http://www.esds.ac.uk/international/support/user_guides/oecd/sti_manual.pdf

Perron, P. (1989). THE GREAT CRASH, THE OIL PRICE SHOCK, AND THE UNIT-ROOT HYPOTHESIS. *Econometrica, 57*(6), 1361-1401. doi: 10.2307/1913712

Perron, P., & Yabu, T. (2009a). Estimating deterministic trends with an integrated or stationary noise component. *Journal of Econometrics, 151*(1), 56-69. doi: 10.1016/j.jeconom.2009.03.011

Perron, P., & Yabu, T. (2009b). Testing for Shifts in Trend With an Integrated or Stationary Noise Component. *Journal of Business & Economic Statistics, 27*(3), 369-396. doi: 10.1198/jbes.2009.07268

Pollitt, Christopher. (1993). *Managerialism and the public services: cuts or cultural change in the 1990s?* (2nd ed.). Cambridge, MA: Blackwell Business.

Quandt, R. E. (1960). TESTS OF THE HYPOTHESIS THAT A LINEAR-REGRESSION SYSTEM OBEYS 2 SEPARATE REGIMES. *Journal of the American Statistical Association, 55*(290), 324-330. doi: 10.2307/2281745

Skoie, H. (1996). Basic research--A new funding climate? *Science and Public Policy, 23*(2), 66-75.
Stephan, Paula E. (1996). The economics of science. *Journal of Economic Literature, 34*(3), 1199-1235.
Thomson Reuters. (2013). ISI. Retrieved May 21, 2013, from http://thomsonreuters.com/products_services/science/science_products/a-z/isi/
Van der Meulen, Barend. (1998). Science policies as principal–agent games: Institutionalization and path dependency in the relation between government and science. *Research Policy, 27*(4), 397-414. doi: 10.1016/s0048-7333(98)00049-3
Vogelsang, T. J., & Franses, P. H. (2005). Testing for common deterministic trend slopes. *Journal of Econometrics, 126*(1), 1-24. doi: 10.1016/j.jeconom.2004.02.004

Appendix 1. The search for lag length and polynomial degree

For lag length selection, we take a maximum seven-year lag and fit eight different models with different lag lengths (i.e., 0, 1, …, 7), and then use the Akaike Information Criteria (AIC), the Bayesian information criterion (BIC), and the Bozdogan Index of Information Complexity (ICOMP) for model selection, that is, the lag length minimizes these information criteria is selected. Crespi and Geuna (2008) took a maximum lag of ten years and evaluated eleven models, and their results suggested the optional order to be six. Given this previous finding and the fact that their cross-country dataset is larger than our individual country dataset, we start with a maximum lag of seven instead of ten years. Results are reported in *Table 5*. Both AIC and ICOMP suggest a six-year lag, while BIC suggest a 0-year lag. Therefore, we adopt a six-year lag, which is the same as Crespi and Geuna (2008).

After the lag order is chosen, a polynomial degree is selected for fitting an Almon model. We start with a 5^{th} degree function and then sequentially reduce the degree and test the reduction. Results are reported in *Table 6*. We do not reject the reduction from 5^{th} to 4^{th}, …, and from 2^{nd} to 1^{st}, but not lower. Therefore a linear function (1^{st} degree polynomial) is chosen to characterize the structure of lag weights. Insignificant χ^2 statistics (*Table 1*) suggest that these constraints are not rejected, that is, they are valid.

Table 6. Backward elimination F-tests

	F
0	4.723**
1-0	10.755***
2-1	1.358
3-2	2.012
4-3	0.126
5-4	0.282

(*) significant at 10%; (**) significant at 5%; (***) significant at 1%.

Table 5. Unrestricted polynomial distributed lag (PDL) models

	Lag 7	Lag 6	Lag 5	Lag 4	Lag 3	Lag 2	Lag 1	Lag 0
	-0.647	-0.693			-0.614	-0.604		-0.643
H	[0.296]**	[0.271]**	-0.586 [0.296]*	-0.605 [0.290]*	[0.289]**	[0.287]**	-0.411 [0.276]	[0.200]***
t-1	-0.054 [0.446]	0.009 [0.414]	0.028 [0.460]	0.020 [0.454]	0.119 [0.440]	0.240 [0.415]	-0.318 [0.265]	
t-2	0.100 [0.523]	0.086 [0.507]	-0.027 [0.561]	0.036 [0.546]	-0.207 [0.480]	-0.528 [0.312]		
t-3	-0.698 [0.554]	-0.717 [0.537]	-0.522 [0.589]	-0.688 [0.536]	-0.283 [0.321]			
t-4	0.376 [0.524]	0.410 [0.505]	0.042 [0.527]	0.329 [0.348]				
t-5	-0.341 [0.482]	-0.422 [0.437]	0.248 [0.337]					
t-6	0.478 [0.456]	0.637 [0.299]*						
t-7	0.151 [0.321]							
	0.071	0.073	0.077	0.080	0.083	0.079	0.073	0.069
Year	[0.013]***	[0.011]***	[0.012]***	[0.011]***	[0.011]***	[0.010]***	[0.009]***	[0.009]***
R2 adj	0.977	0.979	0.974	0.974	0.975	0.975	0.972	0.972
Obs. #	23	23	23	23	23	23	23	23
AIC	-71.118	-72.728	-68.262	-69.451	-70.199	-71.172	-69.771	-70.096
BIC	-58.627	-61.373	-58.043	-60.367	-62.251	-64.359	-64.093	-65.554
ICOMP	-66.809	-67.892	-62.418	-62.943	-63.179	-63.508	-61.641	-60.904

Standard errors reported in brackets. (*) significant at 10%; (**) significant at 5%; (***) significant at 1%.