Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Replication and Robustness Analysis of 'Energy and Economic Growth in the USA: a Multivariate Approach'

CAMA Working Paper 18/2018
April 2018

Stephan B. Bruns
Department of Economics, University of Göttingen, Germany

Johannes König
Department of Economics and INCHER, University of Kassel, Germany

David I. Stern
Crawford School of Public Policy, ANU

Abstract

We replicate Stern (1993, Energy Economics), who argues and empirically demonstrates that it is necessary (i) to use quality-adjusted energy use and (ii) to include capital and labor as control variables in order to find Granger causality from energy use to GDP. Though we could not access the original dataset, we can verify the main original inferences using data that are as close as possible to the original. We analyze the robustness of the original findings to alternative definitions of variables, model specifications, and estimation approach for both the (almost) original time span (1949-1990) and an extended time span (1949-2015). p-values tend to be substantially smaller if energy use is quality adjusted rather than measured by total joules and if capital is included. Including labor has mixed results. These findings tend to largely support Stern's (1993) two main conclusions and emphasize the importance of accounting for changes in the energy mix in time series modeling of the energy-GDP relationship and controlling for other factors of production. We also discuss how the inclusion of the original author in designing the replication study using a pre-analysis plan can help to counterbalance the incentive of replicating authors to disconfirm major findings of the original article to increase the probability of getting published.

Keywords

Replication, robustness analysis, sensitivity analysis, energy, GDP, Divisia index, Granger causality

JEL Classification

Q43, C32, C52

Address for correspondence:

(E) cama.admin@anu.edu.au

ISSN 2206-0332

Replication and Robustness Analysis of 'Energy and Economic Growth in the USA: a Multivariate Approach'

Stephan B. Bruns*

Department of Economics, University of Göttingen, Humboldtallee 3, 37073 Göttingen, Germany. stephan.bruns@uni-goettingen.de. Phone: +49-551-39-21391.

Johannes König

Department of Economics and INCHER, University of Kassel, Mönchebergstr. 17, 34109 Kassel, Germany. koenig@uni-kassel.de. Phone: +49 561 804-2709.

David I. Stern

Crawford School of Public Policy, The Australian National University, 132 Lennox Crossing, Acton, ACT 2601, Australia. david.stern@anu.edu.au. Phone: +61-2-6125-0176.

22 April 2018

Abstract

We replicate Stern (1993, *Energy Economics*), who argues and empirically demonstrates that it is necessary (i) to use quality-adjusted energy use and (ii) to include capital and labor as control variables in order to find Granger causality from energy use to GDP. Though we could not access the original dataset, we can verify the main original inferences using data that are as close as possible to the original. We analyze the robustness of the original findings to alternative definitions of variables, model specifications, and estimation approach for both the (almost) original time span (1949-1990) and an extended time span (1949-2015). *p*-values tend to be substantially smaller if energy use is quality adjusted rather than measured by total joules and if capital is included. Including labor has mixed results. These findings tend to largely support Stern's (1993) two main conclusions and emphasize the importance of accounting for changes in the energy mix in time series modeling of the energy-GDP relationship and controlling for other factors of production. We also discuss how the inclusion of the original author in designing the replication study using a pre-analysis plan can help to counterbalance the incentive of replicating authors to disconfirm major findings of the original article to increase the probability of getting published.

Keywords: Replication, robustness analysis, sensitivity analysis, energy, GDP, Divisia index, Granger causality

JEL Codes: Q43, C32, C52

Acknowledgements: We thank the Australian Research Council for funding under Discovery Project (DP160100756) "Energy Efficiency Innovation, Diffusion and the Rebound Effect" and three anonymous reviewers for their helpful comments.

* Corresponding author: stephan.bruns@uni-goettingen.de

1. Introduction

There is a large literature that analyzes the role of energy use in economic growth and a very large number of studies using time series analysis, especially Granger causality and cointegration tests (Bruns et al., 2014). Despite the importance of the relation between energy use and GDP for, among others, climate policy and development strategy, there is little consensus about their relationship and meta-analyses of the time series literature reveal that many of the published findings are likely to be spuriously statistically significant (Bruns and Stern, in press; Bruns et al., 2014). In this study, we revisit the early analysis by Stern (1993), which was highly influential in shaping the literature on Granger causality between energy use and GDP as indicated by more than 750 citations in Google scholar.[1] We replicate its findings and conduct extensive robustness analyses by updating the time span of the analysis, applying an alternative estimation approach, and considering many alternative model specifications including alternative definitions of variables. Our analysis tends to support Stern's (1993) two main conclusions: that in order to find Granger causality from energy use to GDP it is necessary to adjust energy use for quality, and to include capital, and to a lesser extent labor, as control variables.

The original study (Stern, 1993) analyzes the role of energy use in generating economic output by comparing Granger causality tests between energy use and GDP with Granger causality tests between quality-adjusted energy use and GDP. It also investigates the effect of controlling for capital and labor inputs by comparing Granger causality tests from bivariate and multivariate vector autoregressions (VARs). Quality adjustment is carried out using a Divisia index of energy volume. The index computes the annual change in the index as the cost-share weighted sum of the changes in the logarithms of each individual energy input. The cost shares are averaged over two periods. This means that growth in energy carriers with larger shares of total cost contribute more to the growth of the index but that changes in prices have no effect on the measured quantity index. For example, growth of a high-priced energy carrier such as electricity will have more effect on the energy volume index than on the simple sum of total joules. The assumption is that the high price of electricity is associated with a high marginal product and that the quality-adjusted index better reflects the contribution of energy to production than the simple total number of joules (Stern, 2010). Stern (1993) measures quality-adjusted energy use by a Divisia index of final energy use, while simple energy use is measured by primary energy consumption.

The original study finds that (i) if energy use is quality-adjusted using the Divisia Index, then quality-adjusted energy use Granger-causes GDP in a VAR with capital and labor as control variables while this is not the case if energy use is unadjusted. Stern (1993) argues that quality adjustment of energy use reduces measurement error by taking the omitted quality dimension of energy use into account. Moreover, (ii) bivariate tests do not show Granger causality between quality-adjusted energy use and

[1] Retrieved on 4.04.2018

GDP indicating that capital and labor are important control variables as would be expected if the VAR models a production function relationship.

As Dewald et al. (1986) already pointed out, replications are likely to get published more easily if they disconfirm major findings of the original article. In fact, surveyed editors and co-editors of 11 top economics journals responded that they all would accept a replication that disconfirms the findings of the original study, while only 29% responded that they would also accept a confirming replication (Galiani et al., 2017). This view is also consistent with 78% of published replications disconfirming at least one major finding of the original article (Duvendack et al., 2015). To increase the probability of getting published, replicating authors might try to cast doubt on the reliability of the original findings by conducting the replication in a way that maximizes the probability of disconfirming at least some of the original findings (Galiani et al., 2017). For example, replicating authors may search for some findings that are not robust and then sell this as a failure to be replicated. Clemens (2017) shows that, for 22 of 35 prominent critiques of published articles, the sampling distributions of the parameter estimates were not the same as those in the original article; classifying these critiques more as robustness analyses rather than replications.

This replication study was designed jointly with the original author before any data were collected or any analysis was conducted. We agreed on a research design that counterbalances the incentive of the replicating authors to disconfirm some original findings and the incentive of the original author to confirm all original findings and then published this research design as a pre-analysis plan (Bruns et al., 2017).[2] The analysis was conducted independently from the original author except for clarification questions regarding data sources. A draft of the analysis was written and sent to the original author to jointly discuss the interpretation of the results obtained by closely following the pre-analysis plan. The original author then contributed to writing the final version of the paper.

The inclusion of the original author ensures that a lack of verification does not result from the inability of the replicating authors to apply the original models to the original data and it ensures that the replication results and the findings from the robustness analyses are adequately discussed in the light of the scope of the original study. Our proposal to include the original author in the design of the replication study is, of course, not a counterproposal to replication by independent authors. Our proposal is a complementary approach that helps to counterbalance incentives and it may help to put overly negative replications by independent authors into perspective. However, this proposal does not address the fact that positive replications have a lower probability of getting published. This can be addressed by an increase in the number of journals that introduce replication sections that also publish positive replications, as *Energy Economics*' new replication section does.

[2] The pre-analysis plan (ID: 20170325AA) was published on 25.03.2017 at EGAP (Evidence in Governance and Politics). The exact link can be found in the references (Bruns et al., 2017).

Following Clemens (2017), we distinguish between replications and robustness analyses. Replications ensure that the sampling distributions of the parameter estimates remain unchanged with respect to the original analysis. Otherwise, the original findings may not be replicated due to a change in the sampling distributions of the parameter estimates. Such a change in the sampling distributions may occur due to the analysis of different populations or different model specifications that represent robustness analyses rather than a replication of the original results. Clemens (2017) further distinguishes between two types of replications: verifications and reproductions. Verifications use exactly the same sample and the same specifications as were used in the original article and verify that the reported findings are correct. Reproductions draw a new sample of the same population and apply the same specifications to the new sample. Reproductions are the typical type of replication in experimental research and verifications are the typical type of replications in observational research where there is often only one sample available. According to these definitions, we attempt to verify the findings reported in Stern (1993).

We also conduct extensive robustness analyses of the findings reported in Stern (1993). Clemens (2017) distinguishes between two types of robustness analysis: reanalysis and extension. Reanalysis refers to applying alternative specifications to the original sample. Extension refers to applying the original methods to a new sample from a different population. In the time series context, this might include a longer sample of the same time series variable where the data-generating process may differ.

We update the data used in Stern (1993) from 1947-1990 to 1949-2015 and we include new variables and alternative definitions of the original variables. Based on this new data set, we first reanalyze the (almost) original time span (1949-1990) by estimating a large number of VARs with varying control variables and alternative definitions of energy use, labor, and capital. We also consider a different approach to testing that is more suited to deal with non-stationarity in the time series compared to the approach used in the original study. Second, we extend Stern (1993) by applying the original VAR specification with the original testing technique to 1949-2015. Third, we combine the re-analyses with the extension by also applying the extensive robustness checks that were used for the time span 1949-1990 to the updated time span 1949-2015. Finally, we use meta-regressions to analyze how the use of quality-adjusted energy use and capital and labor as control variables influences the *p*-values generated by the large number of VARs considered in the robustness analysis.

Though we could not access the original dataset, we do verify the main original inferences using data that are as close as possible to the original. In the reanalysis and extension, we find that Granger causality tests of the effect of energy use on GDP that use quality-adjusted energy use do result in systematically smaller *p*-values compared to those that just use total joules. Including capital in the VAR also results in systematically smaller *p*-values though including labor has mixed results. These findings provide support for Stern's (1993) two main conclusions.

The next section verifies the reported results of Stern (1993). Section 3 implements the robustness analyses, Section 4 presents results from the meta-regression analysis, and Section 5 summarizes and concludes.

2. Verification

We start with a verification of the reported findings in Stern (1993), that is, we aim to apply exactly the same VARs to the data that were used in the original article. As outlined in the pre-analysis plan, the exact data used in the original article are not available anymore. David Stern updated the spreadsheet used in Stern (1993) with new data as it became available rather than archiving a copy as used in that paper. The earliest available version of the data used in Stern (1993) is the data that was used in Stern (2000). See Stern (2000) for a description of the differences between the datasets.[3]

We report the original findings of Stern (1993) together with our replication results for the multivariate models including labor and capital in Table 1 and for the bivariate models in Table 2. We focus here on the Granger causality tests regarding energy use and GDP. Full results can be found in Appendix B. All variables are in log levels. For both Table 1 and Table 2, Panel A presents the original and replicated results for VARs with unadjusted primary energy use and Panel B presents the results for VAR models with quality-adjusted final energy use.

The verifications are reported in Columns I and II (Column III to XI are discussed in Section 3) and show that the original and replicated findings are qualitatively the same. Whenever an original Granger causality test was significant at the 0.1 level, this is also true for the replicated finding. For the bivariate models, the test statistics remain very similar. This indicates that changes in the test statistics observed for the multivariate models may stem from data revisions of the control variables capital and labor or joint data revisions of the control variables and energy use and GDP.

We conclude that we can largely verify the results reported in Stern (1993). Quality-adjusted energy use Granger-causes GDP if capital and labor are included as control variables while we do not find Granger causality from unadjusted energy use to GDP in the multivariate analysis nor from both types of energy

[3] Though Stern (2000) reports on results for the period 1948-1994, Stern actually prepared a time series on all variables starting in 1947, which we have used in this paper. The following major changes or improvements were made in the data set used by Stern (2000) compared to Stern (1993):

Labor is measured in hours worked by full-time and part-time employees in domestic industries in Stern (2000). However, the data set also included full-time equivalent (FTE) employment as used in Stern (1993) and, so, we use FTE employment in the verification.

GDP is measured in 1987 prices rather than 1982 prices.

Capital was not adjusted for utilization. However, the unemployment rate was included in the data set and so we have adjusted capital for utilization using the unemployment rate as in Stern (1993).

Energy is improved by expanded reporting of non-utility production of electricity and renewable energy sources and by new data on biomass use etc. Fossil fuel prices for the calculation of the Divisia index were improved by using the expenditure data reported in the US Energy Information Administration's *Annual Energy Review* to obtain better estimates of actual final use fuel prices for oil, natural gas, and coal.

use to GDP in the bivariate analyses. We cannot verify the exact values of the test statistics due to data revisions, but our findings confirm the inferences made in the original study. However, note that the Granger causality test from unadjusted primary energy use to GDP is almost statistically significant at the 0.1 level if the revised data from Stern (2000) is used.

Table 1: Multivariate models with capital and labor

	Original estimation approach					Toda-Yamamoto procedure					
Source or data	Stern (1993) (1947-1990)	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Panel A: Primary energy use											
Energy causes GDP	0.5850	2.2994	3.4284	1.2058	0.1473	0.5987	5.7514	4.4733	4.2997	3.2691	5.706
	(0.5628)	(0.1162)	(0.0452)	(0.314)	(0.8634)	(0.5532)	(0.1244)	(0.2147)	(0.2309)	(0.3520)	(0.1268)
GDP causes energy	9.0908	8.0747	4.4233	6.6703	3.9209	2.523	6.6263	9.7513	9.7044	9.6279	6.7558
	(0.0007)	(0.0014)	(0.0204)	(0.0041)	(0.0255)	(0.0898)	(0.0848)	(0.0208)	(0.0213)	(0.0220)	(0.0801)
Panel B: Quality-adjusted final energy use											
Energy causes GDP	3.1902	5.0755	2.0688	3.2165	1.5369	2.5347	8.5235	5.9886	9.9734	6.1342	9.3671
	(0.0319)	(0.0044)	(0.1212)	(0.0356)	(0.2073)	(0.0538)	(0.0363)	(0.1122)	(0.0188)	(0.1053)	(0.0248)
GDP causes energy	0.8458	1.0851	1.4793	1.4183	3.1269	1.6284	4.2081	5.5769	5.4936	4.5703	3.3684
	(0.5106)	(0.3870)	(0.2441)	(0.2661)	(0.0234)	(0.1845)	(0.2399)	(0.1341)	(0.139)	(0.2061)	(0.3382)

Notes: Results in Column I are taken from Tables 6 and 10 in Stern (1993). Original estimation approach refers to F-tests based on VARs in log levels as in Stern (1993). Lag lengths for Columns I to VI in Panel A are 2 and in Panel B are 4 as in Stern (1993). Toda-Yamamoto procedure refers to χ^2-tests based on VARs augmented by one lag. For both Panels A and B, Columns VII to XI use 3(+1) lags as specified by using the AIC using a maximum potential lag length of 3. A maximum lag length of 3 was chosen to ensure sufficient degrees of freedom (see Section 3.2). *p*-values in parentheses. Diagnostic tests are available in the Appendix B.

Table 2: Bivariate models

	Original estimation approach					Toda-Yamamoto procedure					
Source or data	Stern (1993) (1947-1990)	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Panel A: Primary energy use											
Energy causes GDP	0.8328 (0.4428)	0.9270 (0.4047)	0.3455 (0.7103)	0.9998 (0.3788)	0.0284 (0.9720)	0.6359 (0.5332)	1.6408 (0.2002)	0.0136 (0.9073)	0.4973 (0.4807)	0.0543 (0.8157)	0.0442 (0.8335)
GDP causes energy	0.3421 (0.7125)	0.2854 (0.7534)	0.5316 (0.5923)	3.2149 (0.053)	0.0925 (0.9118)	1.083 (0.3454)	0.0089 (0.9248)	0.7915 (0.3737)	0.3601 (0.5484)	0.1601 (0.6890)	0.0000 (0.9939)
Panel B: Quality-adjusted final energy use											
Energy causes GDP	0.9657 (0.4402)	0.8339 (0.5140)	0.9703 (0.4388)	1.1571 (0.3515)	1.0584 (0.3878)	1.2801 (0.29)	3.3210 (0.3447)	0.8577 (0.3544)	0.1855 (0.6667)	1.2607 (0.2615)	0.9726 (0.324)
GDP causes energy	0.7154 (0.5878)	0.6031 (0.6633)	1.4782 (0.2344)	2.3517 (0.0793)	0.7880 (0.5380)	1.5271 (0.2082)	2.2964 (0.5132)	2.5073 (0.1133)	0.0722 (0.7881)	2.0772 (0.1495)	0.6865 (0.4074)

Notes: Results in Column I are taken from Table 7 and 11 in Stern (1993), respectively. Original estimation approach refers to F-tests based on VARs in log levels as in Stern (1993). Lag lengths for Columns I to VI in Panel A are 2 and in Panel B are 4 as in Stern (1993). Toda-Yamamoto procedure refers to χ^2-tests based on VARs augmented by one lag. For both Panel A and B, Column VII to XI have 1(+1) lags specified by using the AIC with a maximum potential lag length of 3. The only exception is in Column VII (Panel B), which uses 3(+1) lags. A maximum lag length of 3 was chosen to ensure sufficient degrees of freedom (see Section 3.2). *p*-values in parentheses. Diagnostic tests are available in the Appendix B.

3. Robustness analysis

3.1 Data

We explore the robustness of the results presented in Stern (1993) to an updated time span, to an alternative estimation technique, and to various alternative specifications of the VARs including alternative definitions of the original variables. To this end, we built a new data set based on the original data.

The main sources of data that were used in Stern (1993) are still available and are updated by the same agencies on a yearly/monthly base, but data revisions mean the current series have different values for the 1949-1990 period than those used by Stern (1993) or Stern (2000). Data for GDP and the price index

of GDP are taken from the Bureau of Economic Analysis (BEA, 2017a; BEA, 2017c). Table A1 in Appendix A gives a detailed overview of the data and their sources.

Stern (1993) uses private and government capital but excludes residential capital. We also collect data on capital including residential capital to explore robustness to this alternative definition of capital (BEA, 2017a; BEA, 2017c). Stern (1993) used (1 – unemployment rate) as a proxy of the utilization rate of capital. Data for the unemployment rate were obtained from the Bureau of Labor Statistics (DOL, 2017). We outlined in the pre-analysis plan that we would collect data on the actual utilization rate to explore robustness, but this data is not available from 1949 for the whole economy. Therefore, we deviate from the pre-analysis plan and, instead, explore robustness to the use of capital as a control variable with and without multiplication by the proxy utilization rate. Hence, each of the two alternative definitions of capital can be multiplied either with the proxy utilization rate or not resulting in four different measures of capital.

Stern (1993) measured labor using full-time equivalent (FTE) employment. We also collect data on hours worked as an alternative definition of this control variable, resulting in two alternative measures of labor. Data for labor are taken from the Bureau of Economic Analysis (BEA, 2017b).

Energy related data are obtained from the Energy Information Administration (EIA, 2017).[4] Stern (1993) analyzes unadjusted primary energy use and quality-adjusted final energy use. We also assembled data on quality-adjusted primary energy use and unadjusted final energy use to explore robustness to these alternative definitions. We base our data collection on the most recent version of the *Monthly Energy Review* (March 2017) (EIA, 2017) where possible. The *Monthly Energy Review* reports time series only from 1949. The data used in Stern (1993) goes back to 1947 and data for the two missing years is based on historical data reported in the *Historical Statistics of the United States*. While in principle we could extrapolate the data backwards using the *Historical Statistics of the United States*, hours of employment are available only from 1948 and cannot be extrapolated. Instead of analyzing 1948-2015 by extrapolating one year back using data from the *Historical Statistics*, we decided to deviate here slightly from the pre-analysis plan and analyze the period 1949-2015, which allows us to use the data directly reported in the *Monthly Energy Review* in March 2017.

We implement quality adjustment using the Divisia index, which requires data on energy prices for all energy types (e.g. electricity, oil, natural gas, biomass). However, the final energy prices are first reported in the *Monthly Energy Review* (and the now discontinued *Annual Energy Review*) from the end of the 60's. We extrapolate final energy prices backwards by using the growth rates of the final energy prices in the data used in Stern (2000).[5] This permits us to obtain time series of energy prices that range from 1949 to 2015. We quality adjust primary energy use using primary energy prices and we quality

[4] The only exception is the price of lumber, which is obtained from the *Historical Statistics of the United States*.
[5] Stern (2000) used the growth rates of production prices to extrapolate backwards the end-user prices implied by the expenditure data.

adjust final energy use using final energy prices. Note that Stern (1993) uses 60% of the price of coal to proxy the price of biomass and the 'other' category. However, we realized that prices for these categories actually tend to be greater than 60% of the price of coal. For the reanalysis, we, therefore, use primary energy prices for these two categories to quality adjust final energy use. Overall, we consider four different types of energy use: unadjusted primary and final energy use and quality-adjusted primary and final energy use.

Finally, in some of the VARs used in our reanalysis, we include energy prices as a control variable. In these specifications, we use the energy price that matches the definition of energy use used in the VAR. We calculate simple energy prices for primary and final energy use by dividing overall expenditures for final or primary energy use by energy use of these categories. We also calculate quality-adjusted primary and final energy prices using the Divisia index.

Figure 1 gives an overview of the original and updated data. The original and updated time series are similar, however, due to data revisions, there are differences between these time series. Revisions to GDP have resulted in a higher measured rate of economic growth than shown in the original series. Similarly, capital grows faster in the 1980s than in the original data. Including residential capital does not make a very large difference – growth in capital is mainly affected in the period following the Great Recession of 2008-2009 due to a slowdown in housing construction following the bursting of the housing price bubble. Unadjusted final energy use grows more slowly than unadjusted primary energy use because of the increasing share of electricity in final energy use over time. By contrast, quality-adjusted final energy use grows more rapidly than quality-adjusted primary energy use because of the same shift to electricity. Updated FTE employment data show a somewhat faster rate of employment growth in all decades. Hours grow more slowly than FTE, as the hours worked by a full-time employee declined over time. Finally, though the overall patterns in energy prices are similar, the various indices differ quite radically. Quality-adjusted final energy prices rise by much less than unadjusted final energy prices because the quantity of quality-adjusted final energy use increases by more than unadjusted final energy use and the product of price and quantity must be the same for both quality-adjusted and unadjusted energy use.

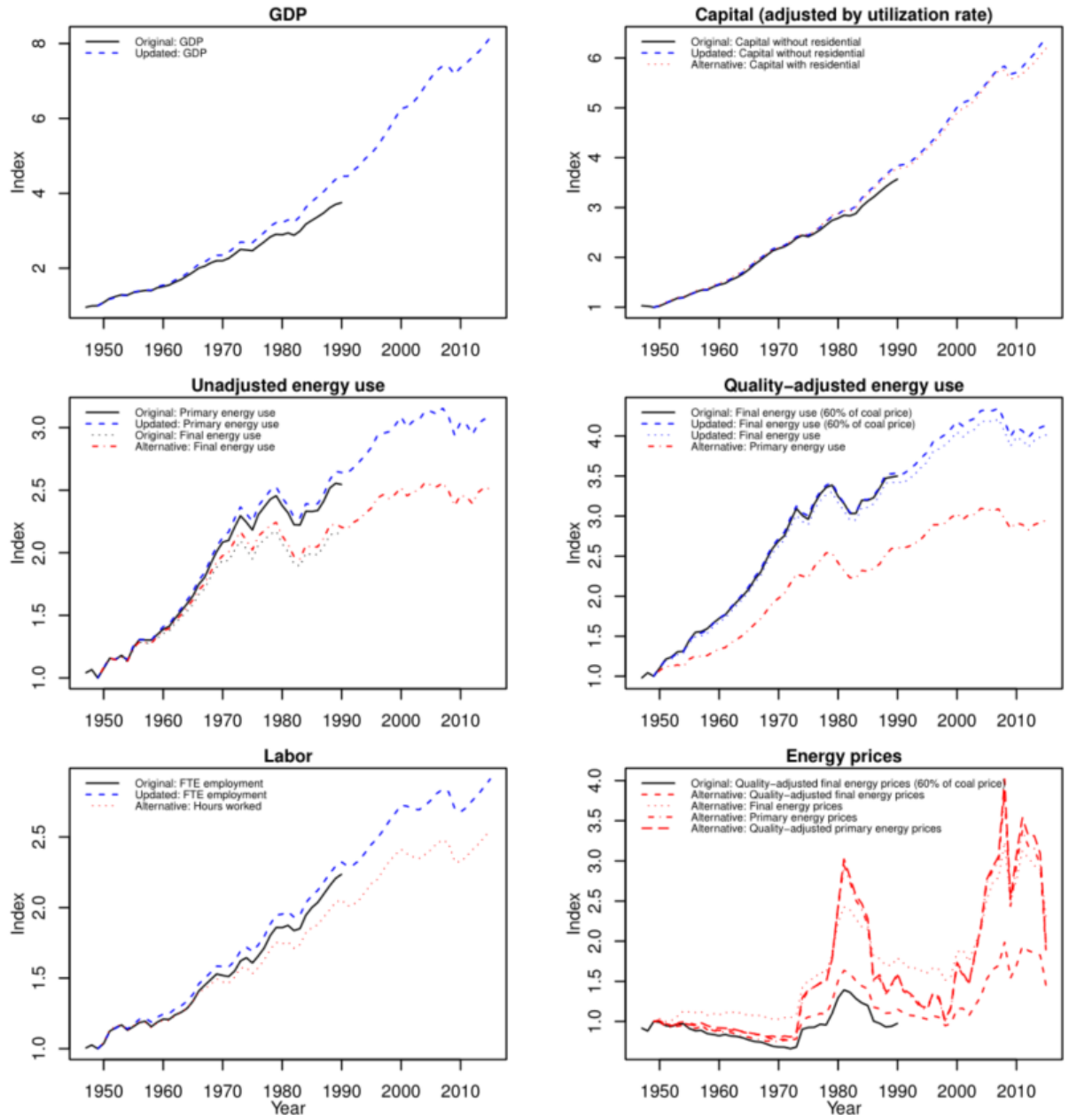

Figure 1: Original (black) refers to the original Stern (2000) data. Unadjusted final energy use and energy price variables were not used in the analysis of Stern (1993) but are included in the Stern (2000) data and reported here to assess data revisions. Updated (blue) refers to updated time series for the variables that were actually used in Stern (1993). Alternative (red) refers to time series that were not considered in Stern (1993) but are used here to estimate alternative model specifications. The exact definition of each time series is given in Table A1 in Appendix A.

3.2. Reanalysis methods

After verifying the original findings, we continue with robustness analyses by means of reanalysis of the (almost) original period 1949-1990, analysis of the extended period 1949-2015 and reanalysis of the extended period. Our reanalysis includes alternative model specifications and alternative definitions of

the variables as well as an alternative testing approach. Each of these changes in the analysis may alter the sampling distributions of the parameter estimates and cannot be considered as a direct replication of the original findings.

We conduct unit root tests for all variables and they suggest that all the time series considered have a unit root.[6] This suggests that the Granger causality tests in levels used by Stern (1993) are biased, as they do not take into account the non-stationarity of the time series (Toda and Phillips, 1993). We explicitly address the non-stationarity of the time series by analyzing the data using the modified Granger Causality procedure suggested by Toda and Yamamoto (1995). They show that if a VAR in levels is augmented by the number of lags equal to the maximum degree of integration of the analyzed time series, then a Wald test that ignores these additional lags is asymptotically χ^2-distributed. This holds irrespective of whether the time series are integrated or cointegrated.

As an alternative to the procedure suggested by Toda and Yamamoto (1995), one could test for cointegration and then estimate a Vector Error Correction Model (VECM) if the analyzed time series are cointegrated or a VAR in first differences if they are not cointegrated. However, this pre-testing can increase the rate of false-positive findings of Granger causality (Clarke and Mirza, 2006). In the pre-analysis plan, we stated that we would use Granger causality tests based on VARs in first differences to corroborate the results of the Toda-Yamamoto procedure. Two anonymous reviewers suggested that we remove the analysis in first differences and consider structural breaks instead as there is some indication that the time series analyzed here are cointegrated (e.g. Stern, 2000; Oh and Lee, 2004; Shahiduzzaman and Alam, 2012) and this would bias the analysis in first differences. We follow these suggestions and moved the analysis in first differences, which largely confirms the findings obtained by the Toda-Yamamoto procedure, to the Appendix B. Instead we added an analysis that accounts for structural breaks.

We explore robustness of the key results to various alternative variable definitions and specifications of the underlying VARs. The stylized model that we use for this robustness analysis is given by

$$\begin{bmatrix} y_t \\ e_t^{(k)} \\ \boldsymbol{z}_t^{(k)} \end{bmatrix} = \begin{bmatrix} \alpha_1^{(k)} \\ \alpha_2^{(k)} \\ \boldsymbol{\alpha}_3^{(k)} \end{bmatrix} + \sum_{i=1}^{m^{(k)}} \begin{bmatrix} \beta_{11,i}^{(k)} & \beta_{12,i}^{(k)} & \boldsymbol{\beta}_{13,i}^{(k)} \\ \beta_{21,i}^{(k)} & \beta_{22,i}^{(k)} & \boldsymbol{\beta}_{23,i}^{(k)} \\ \boldsymbol{\beta}_{31,i}^{(k)} & \boldsymbol{\beta}_{32,i}^{(k)} & \boldsymbol{\beta}_{33,i}^{(k)} \end{bmatrix} \begin{bmatrix} y_{t-i} \\ e_{t-i}^{(k)} \\ \boldsymbol{z}_{t-i}^{(k)} \end{bmatrix} + \begin{bmatrix} \epsilon_1^{(k)} \\ \epsilon_2^{(k)} \\ \boldsymbol{\epsilon}_3^{(k)} \end{bmatrix} \tag{1}$$

where $t = 1, \dots, T$ is the index for time, $m^{(k)}$ represents the lag length, $k = 1, \dots, K$ is the model index that will be discussed below in more detail and vectors are in bold. GDP is given by y and is the same for all k. The energy variable, $e^{(k)}$, represents either primary or final energy use that is either unadjusted or quality-adjusted depending on k. The vector of control variables, $\boldsymbol{z}^{(k)}$, including its dimension depends on k as well. The model in (1) is augmented by one lag, as the maximum order of integration

[6] See the Online Appendix for results.

is one. For the estimation with structural breaks, we allow for break points in the constant for the two oil crises (1973 and 1979) and the global financial crisis (2008). All variables are in logs.

We focus here on the robustness of Granger causality tests from energy use to GDP and *vice versa*. The null hypothesis of Granger non-causality[7] from energy use to GDP is given by

$$H_0: \beta_{12,1}^{(k)} = \cdots = \beta_{12,m^{(k)}}^{(k)} = 0 \tag{2}$$

and the null hypothesis of Granger non-causality from GDP to energy use is given by

$$H_0: \beta_{21,1}^{(k)} = \cdots = \beta_{21,m^{(k)}}^{(k)} = 0. \tag{3}$$

While a robustness analysis of effect sizes using confidence intervals would be preferable (e.g. Cumming, 2014), Granger causality tests are joint significance tests of multiple lags in a reduced form VAR model and, thus, effect sizes cannot be analyzed. Our robustness analysis is based on the ideas of Leamer (1983) and Leamer and Leonard (1983).

We explore robustness by summarizing the distribution of *p*-values for four groups of VARs. The first group (Group 1) of VARs investigates uncertainty regarding the definitions of the control variables. The VARs in this group always include capital and labor as control variables and uses quality-adjusted energy use. There are two measures of energy use (primary and final), two measures of capital (with and without residential capital) and each of these two capital measures is either multiplied by the utilization rate or not and there are two measures of labor (FTE employment and hours worked). Thus, the first set of VARs comprises $K = 16$ models.

Group 2 allows us to assess the role of quality adjustment of energy use on the *p*-values of Granger causality tests. These VARs use unadjusted energy use instead of quality-adjusted energy use but are otherwise the same as those in Group 1. Group 2 also comprises 16 VARs.

Group 3 uses quality-adjusted energy use and varies the set of control variables, but capital and labor are never included simultaneously. Comparing Group 1 with Group 3 allows us to assess whether indeed both capital and labor are needed as control variables to find Granger causality from energy use to GDP. The set of considered control variables includes the different definitions of capital, labor and energy prices. However, we restrict the dimension of the VARs to four to ensure that the degrees of freedom do not become too small as Bruns and Stern (in press) show that if the maximum lag length is set in a way that may result in very low degrees of freedom being selected by the information criteria, spuriously significant Granger causality tests can occur.[8] Hence, there can be at most two of the three control

[7] As an anonymous reviewer pointed out, the energy-growth literature oversimplifies Granger causality. Granger causality in a multivariate model is actually more complicated (see Lütkepohl, 2005).

[8] Note that because of the extra lag needed for the Toda-Yamamoto test the degrees of freedom for a VAR with 3 lags and dimension 4 estimated using data from 1949 to 1990 is: (1990-1953)-(4*4+1)=20 and 18 for the case with structural breaks.

variables simultaneously in the VAR, but Group 3 also considers all cases where only one control variable is present or where none of the control variables are included. Overall, the third group of VAR models comprises $K = 28$ models. Group 4 uses unadjusted energy use but is otherwise the same as Group 3.

We conduct the robustness analysis separately with and without structural breaks and separately using the two information criteria AIC and BIC that are used to specify the lag length.

If the null hypothesis of Granger non-causality is true and we observe a series of independent Granger causality tests (e.g. based on repeated random sampling of the same population), we would expect the *p*-values to be uniformly distributed with mean 0.5. However, the Granger causality tests obtained from the reanalysis are dependent, as they are based on largely the same sample and only variable definitions and control variables vary. This implies that we can observe any distribution of *p*-values. For example, *p*-values may scatter widely if alternative VAR specifications affect the Granger causality tests, but they may also center around a specific *p*-value if alternative VAR specifications do not have a strong effect on the Granger causality tests. For example, alternative VAR specifications may affect Granger causality tests due to omitted-variable bias (Lütkepohl, 1982).

A strong test of the robustness of the original conclusions implies that Granger causality tests from energy use to GDP that are based on VARs with quality-adjusted energy use and capital and labor as control variables (Group 1) should result in exclusively statistically significant *p*-values while Group 2 to Group 4 should result in exclusively non-significant *p*-values. In this case, both quality-adjustment of energy use and capital and labor control variables are required to find Granger causality from energy use to GDP irrespective of the various alternative variable definitions. However, even if Stern's (1993) conclusions are correct and robust to the alternative variable definitions and specifications considered in this study, we do not expect the *p*-values to be exclusively statistically significant or non-significant. Each re-estimation may result in a false-positive or false-negative finding due to sampling variability. Hence, we use this "strong robustness" as a benchmark for discussing the results.

An anonymous reviewer suggested to adjust the *p*-values for multiple testing to control the (expected) rate of type I errors. This would allow us to draw inference on Granger causality from single model specifications using a significance level of 0.05 or 0.1 that holds even in the presence of multiple testing. However, this adjustment comes at the costs of a loss of power and makes the robustness analysis overly conservative. We analyze the distributions of *p*-values for each of the four groups introduced above to explore robustness and our aim is not to use these models to draw conclusions based on single *p*-values. Nevertheless, we present results for adjusted *p*-values in the Appendix B to show how this affects the results.

The analysis was conducted in R version 3.3.1 (R Core Team, 2017) and is mainly based on the packages "vars" (Pfaff, 2008) and "car" (Fox and Weisenberg, 2011).

3.3. Reanalysis for the period 1949-1990

Before discussing the reanalysis outlined in Section 3.2, we analyze to what extent updating the data affects the Granger causality tests. This can also be interpreted as a reanalysis, as data revisions may alter the sampling distributions of the parameter estimates. We apply the original models used in Stern (1993) to the new data set.[9] These test statistics can then be compared to the results reported in the original study and those obtained in the verification presented in Section 2. Column III in Table 2 shows that all Granger causality tests remain statistically non-significant for the bivariate models while Column III in Table 1 shows that tests of Granger causality from unadjusted energy use to GDP become statistically significant and the corresponding test for quality-adjusted energy use becomes marginally non-significant. This reverses the conclusions of Stern (1993). If we consider the model with structural breaks in 1973 and 1979, the original inferences are confirmed (Column IV). However, the original estimation approach of Stern (1993) results in biased Granger causality tests due to the non-stationarity of the time series.

Therefore, we also apply the Toda-Yamamoto procedure to the Stern (2000) data and the updated data presented in Columns VII and VIII in Tables 1 and 2. For the Stern (2000) data the original inferences are robust, but for the updated data the Granger causality test from quality-adjusted final energy use to GDP becomes marginally non-significant ($p = 0.1122$). For the VARs with structural breaks, the original inferences are again confirmed (Column IX).

The reanalysis to explore robustness to various alternative variable definitions and model specifications is presented in Figure 2. We report findings for both directions of causality for completeness but focus on Granger causality tests from energy use to GDP as Stern's (1993) conclusions involve only this direction of causality. Panel A in Figure 2 reveals that there is a sharp difference between Granger causality tests from VARs specified using the AIC or from VARs specified using the BIC. AIC-specified VARs result in distributions with smaller *p*-values compared to BIC-specified VARs. The potential source of this difference is the identified lag length. While using the BIC results mostly in one or two lags, using the AIC results mostly in three lags. Kilian and Lütkepohl (2017) recommend using the AIC to avoid underestimation of the lag length in finite samples. However, there is also a danger of overfitting in small samples if the degrees of freedom become few (Bruns and Stern, in press). As discussed in the previous section, we ensure here that there are sufficient degrees of freedom to avoid overfitting and, thus, we put more emphasize on the VARs based on the AIC for this relatively small sample size.

Group 1 explores robustness with respect to alternative variable definitions for VARs with quality-adjusted energy use and capital and labor as control variables. For AIC based VARs, the *p*-values scatter

[9] Note that quality adjustment is based for these estimations on the use of 60% of the coal price for the price of biomass and the 'other' category, exactly as is the case in Stern (1993).

closely around the *p*-value of the original model ($p = 0.1122$). Specifically, 63% (10 out of 16) of the Granger causality tests are statistically significant at the 0.1 level. In addition to the VAR with the original variable definitions, one further model is almost statistically significant and the remaining four models range between 0.39 and 0.49. These *p*-values tend to be smaller than the *p*-values obtained using unadjusted energy use (Group 2). In fact, 81% (13 out of 16) models are statistically non-significant in Group 2. While we do not find robustness in terms of exclusively statistically significant *p*-values for Group 1 and exclusively non-significant *p*-values for Group 2, we find that quality adjustment of energy use tends to reduces *p*-values in VARs with capital and labor.

Group 3 comprises VARs with quality-adjusted energy use where capital and labor are never included together. The range of obtained *p*-values becomes larger, and VARs with capital tend to result in smaller *p*-values. Specifically, 5 out of the 6 VARs that are statistically significant use capital as a control variable. This suggests that capital and labor may not both be needed to find Granger causality from energy use to GDP and that capital may play a more important role. For Group 4, which comprises VARs with unadjusted energy use that never include capital and labor together, only 3 out of 28 VARs result in a statistically significant *p*-value, all of which include capital as a control variable.

VARs specified using the BIC never result in statistically significant Granger causality tests using either quality-adjusted or unadjusted energy use or any set of control variables, but we again find that *p*-values tend to be smaller when we use quality-adjusted energy use and capital as control variables. The results for VARs that allow for structural breaks in 1973 and 1979 are shown in Figure 3. *p*-values tend to be slightly smaller but otherwise the results are largely the same.

The reanalysis for Granger causality from GDP to energy use (Panel B in Figure 2 and 3) shows that most of the *p*-values are smaller than 0.1 for VARs with capital and labor irrespective of whether quality-adjusted or unadjusted energy use is considered. This, however, does not hold for the other VAR specifications.

Overall, we find evidence that data revisions alter the inferences if the original estimation approach is used, which however results in biased Granger causality tests. Using the Toda-Yamamoto procedure shows that the inferences remain largely stable except in the case of the Granger causality test from quality-adjusted final energy use to GDP, which becomes marginally non-significant. Allowing for structural breaks confirms the original inferences for both estimation procedures. The original inferences are not strongly robust with respect to alternative variable definitions, but using quality-adjusted energy use results in systematically smaller *p*-values compared to using unadjusted energy use. The use of capital and labor as control variables also tends to reduce *p*-values and capital seems to be particularly important.

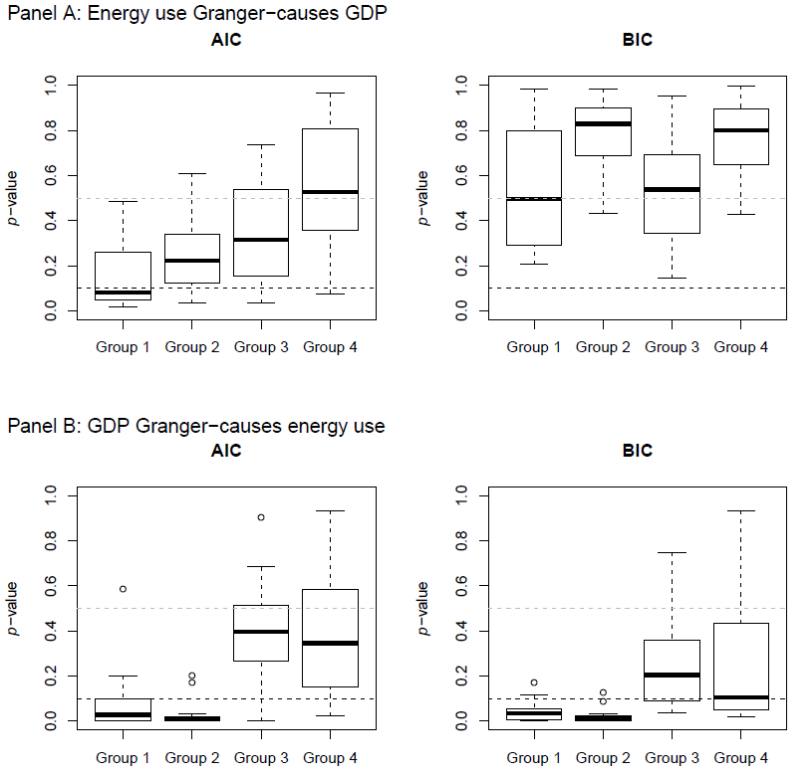

Figure 2. Reanalysis of the (almost) original time span 1949-1990. Boxplots for *p*-values of Granger causality tests are shown separately for VARs specified with AIC and BIC. Group 1 comprises VARs with quality-adjusted energy use and capital and labor. Group 2 is the same as Group 1 but uses unadjusted energy use. Group 3 comprises VARs with quality adjusted energy use but capital and labor are never included simultaneously as control variables. Group 4 is the same as Group 3 but uses unadjusted energy use. Dashed lines represent *p*-values of 0.5 and 0.1. Whiskers of the boxplots extend to the minimum and maximum *p*-value within a multiple of 1.5 times the interquantile range and outliers are depicted by dots. The black solid lines represent means.

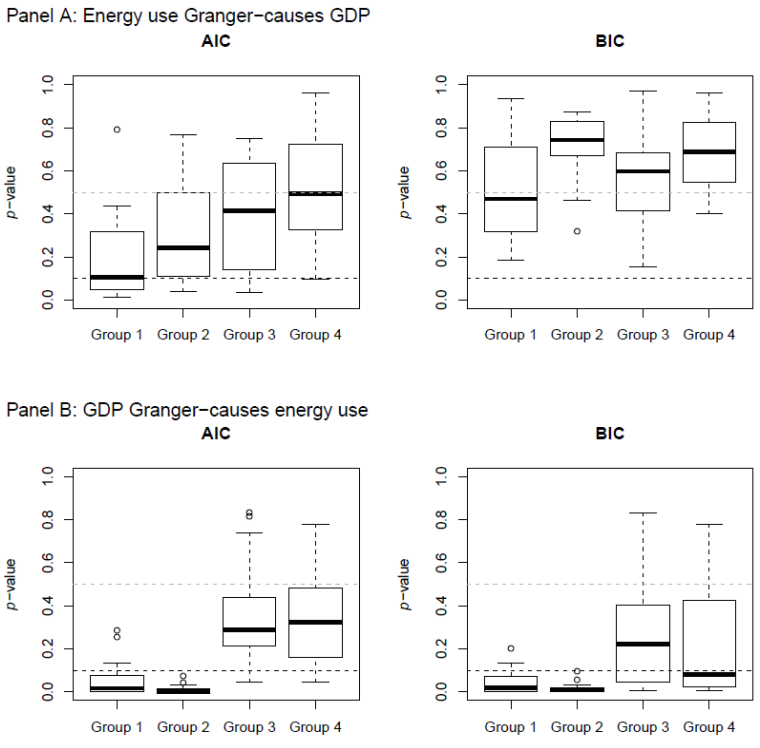

Figure 3. Reanalysis of the (almost) original time span, 1949-1990, with structural breaks in 1973 and 1979. Please see caption of Figure 2 for further details.

3.4. Extension to the period 1949-2015

First, we extend the original analysis to the period 1949-2015 before we carry out the reanalysis for this extended period. If the data-generating process remains stable, the application of the original estimation approach to 1949-2015 could be considered a reproduction with a larger sample size. If the data-generating process changes, this analysis may also be considered an extension. But the previous section suggests that data revisions may alter the inferences and, thus, this analysis may also be considered as a combination of extension and reanalysis or reanalysis with a larger sample size depending on the stability of the data-generating process.

Column V in Tables 1 and 2 report the results for the original estimation approach. For the multivariate model, Granger causality tests from both quality-adjusted and unadjusted energy use to GDP are statistically non-significant. If structural breaks in 1973, 1979, and 2008 are included, the original inferences are confirmed (Columns VI). But note again that these Granger causality tests are biased. Therefore, we apply the Toda-Yamamoto procedure to the period 1949-2015 as well. These results are reported in Columns X. For the multivariate model, the Granger causality test from quality-adjusted final energy use to GDP is marginally non-significant ($p = 0.1053$) but inferences are otherwise consistent with the original findings. If we consider structural breaks, the original inferences are again confirmed (Columns XI). Note that the VARs with primary energy use show signs of non-normality and autocorrelation, but both are resolved if structural breaks are considered.

3.5. Reanalysis for the period 1949-2015

We also apply the reanalysis outlined in Section 3.2 to the updated data for the period 1949-2015. The results are similar to those obtained for the period 1949-1990 and are presented in Figure 4. For Granger causality tests from energy use to GDP (Panel A), the differences between *p*-values obtained for quality-adjusted energy use and unadjusted energy use tend to be larger than those for 1949-1990. Moreover, there is less variation in *p*-values for the well-specified models (Group 1). Results obtained for AIC-specified VARs and BIC-specified VARs look more similar in this larger sample, though differences remain. The BIC selects mostly two lags while the AIC picks mostly three lags.

The results for the VARs with structural breaks in 1973, 1979, and 2008 are presented in Figure 5. Results are similar to those without structural breaks, but *p*-values tend to be generally smaller. A notable difference is that for AIC-specified VARs, Granger causality tests from energy use to GDP are almost exclusively statistically significant if quality-adjusted energy use and capital and labor are used (Group

1) while only two Granger causality tests are statistically significant if unadjusted energy use is used (Group 2).

For Granger causality from GDP to energy use (Panel B in Figure 4 and 5), the *p*-values tend to be less significant for Group 1 and 2 compared to the period 1949-1990 and using the BIC tends to result in smaller *p*-values.

In conclusion, we again find that quality adjustment of energy use results in smaller *p*-values compared to energy use that is unadjusted for Granger causality tests from energy use to GDP. This difference tends to be greater than for the 1949-1990 sample. Furthermore, we find again that capital seems to play an important role in reducing *p*-values.

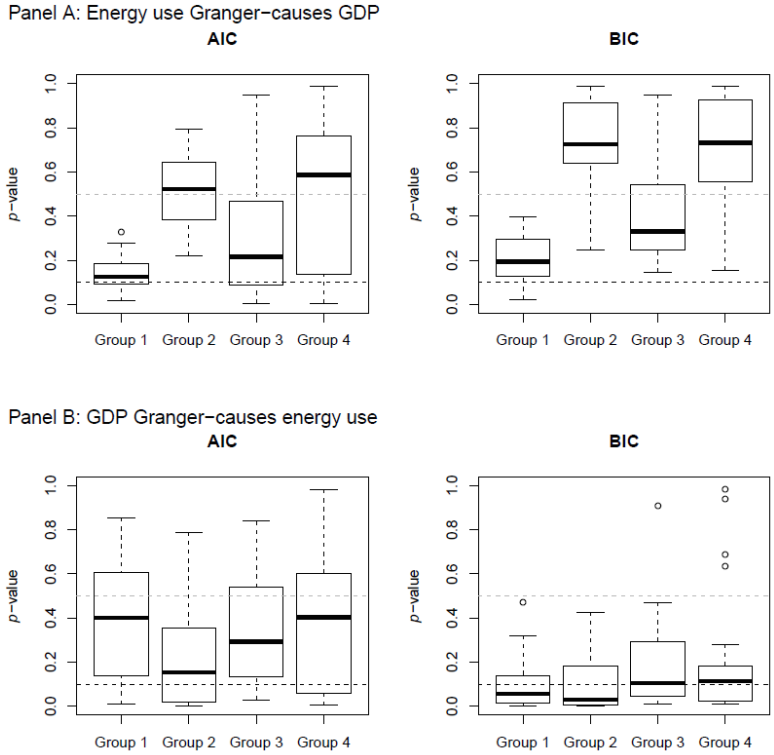

Figure 4. Reanalysis of the extended time span 1949-2015. Please see caption of Figure 2 for further details.

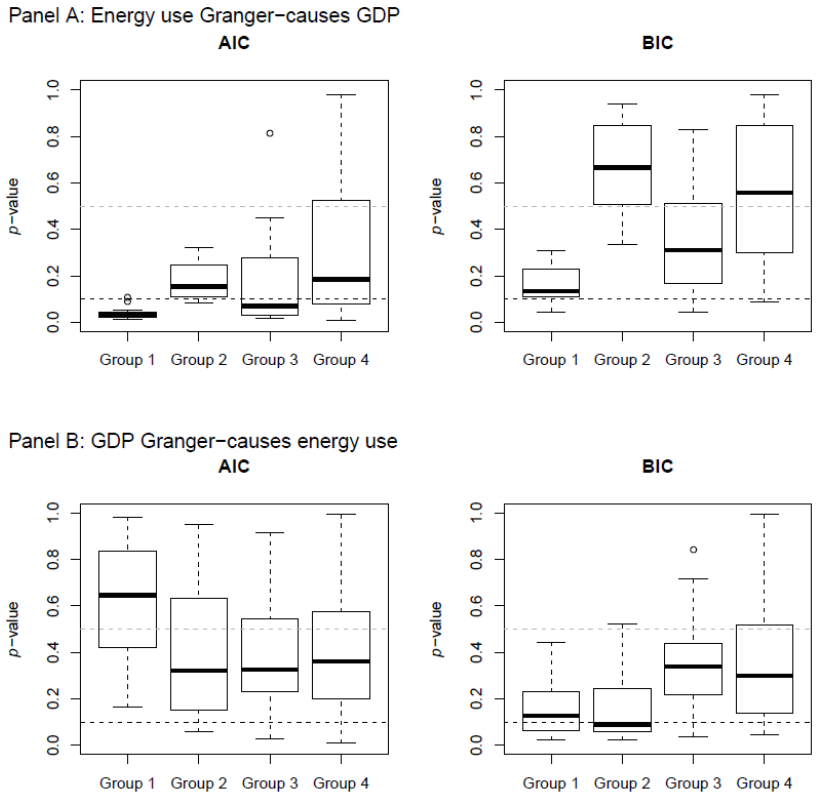

Figure 5. Reanalysis of the extended period 1949-2015, with structural breaks in 1973, 1979, and 2008. Please see caption of Figure 2 for further details.

4. Determinants of the variation in *p*-values

The previous section shows that the original inferences regarding Granger causality from energy use to GDP are largely robust if the original variable definitions and the Toda-Yamamoto procedure are used. The results are not strongly robust to various alternative variable definitions and VAR specifications. But we find that *p*-values tend to be lower if energy use is quality adjusted and if capital and to a lesser extent labor are used as control variables. In this section, we provide a systematic analysis of these tendencies using meta-regressions (see Bruns (2017) for an overview). This section goes beyond the analysis outlined in the pre-analysis plan.

We use probit-transformed *p*-values as the dependent variable (Bruns and Stern, in press; Bruns et al., 2014). This ensures that the *p*-values are standard normally distributed under the null of Granger non-causality resulting in desirable residual properties. We estimate the meta-regressions separately for the two analyzed time periods and for the two directions of causality between energy use and GDP. Each meta-regression is based on the VARs estimated for Group 1 to Group 4 (88 models) where each group is estimated for VARs with and without structural breaks and for VARs specified by the AIC and by the BIC, resulting in *p*-values from 352 different models. Specifically, we estimate

$$z_i = \alpha + \beta_1 C_i + \beta_2 L_i + \beta_3 P_i + \beta_4 Q_i + \beta_5 BIC_i + \beta_6 SB_i + \epsilon_i \quad (4)$$

where $z_i = \Phi^{-1}(p_i)$ are the probit-transformed p-values with $i = 1 \ldots, 352$, C_i is a dummy that is one if p-value i was obtained by using capital as a control variable and zero otherwise, L_i is a dummy that is one if p-value i was obtained by using labor as a control and zero otherwise, P_i is a dummy that is one if p-value i was obtained by using energy price as a control variable and zero otherwise, Q_i is a dummy that is one if p-value i was obtained by using quality-adjusted energy use and zero otherwise, BIC_i is a dummy that is one if p-value i was obtained by using the BIC and zero otherwise, and SB_i is a dummy that is one if p-value i was obtained by taking structural breaks into account and zero otherwise. The benchmark group is thus represented by p-values obtained from bivariate Granger causality tests without considering structural breaks using the AIC to specify the lag length. Smaller values of p_i result in smaller values of z_i and *vice versa*. Note that an important determinant of statistical significance is the number of degrees of freedom. But we do not include the degrees of freedom here as we only have two different sample sizes and we estimate the meta-regression separately for each period. Hence, the variation in degrees of freedom stems exclusively from variation of the VAR specifications.

Columns I and II in Table 3 present the results for tests of whether energy causes GDP. Using the BIC to select the lag length results in less significant Granger causality tests. Quality adjustment of energy use has a negative and significant effect on the z-values, that is, Granger causality tests based on quality-adjusted energy use result in smaller p-values compared to tests based on unadjusted energy use while controlling for many other characteristics of the estimation process. This finding is also in line with the results presented in Figures 2 to 5 and tends to support Stern's (1993) conclusion that quality adjustment is important. Interestingly, the meta-regressions reveal that the use of capital as a control variable results in more significant Granger causality tests from energy use to GDP while the inclusion of labor as a control variable does not seem to matter. These findings tend to partially support Stern's (1993) conclusion that capital and labor are important control variables. An interaction of capital and labor was not statistically significant and is not reported. Taking structural breaks into account matters for the longer period. The effect of using energy prices as a control variable is mixed.

The results for tests of whether GDP causes energy are presented in Columns III and IV. Stern (1993) does not make any claims about this direction of causality. Using labor as a control variable and BIC instead of AIC to select the lag length reduces p-values while the use of quality-adjusted energy use, including energy prices, and considering structural breaks is more mixed.

Table 3: Meta-regression results

	Energy causes GDP		GDP causes energy	
	1949-1990	1949-2015	1949-1990	1949-2015
	I	II	III	IV
Capital	-0.5865***	-0.6389***	-1.1025***	-0.168

	(0.0891)	(0.1008)	(0.1078)	(0.1133)
Labor	-0.0389	-0.1378	-1.5074***	-0.5673***
	(0.0828)	(0.0937)	(0.1002)	(0.1053)
Energy price	0.1856**	-0.1344	0.2423**	0.0902
	(0.0902)	(0.102)	(0.109)	(0.1146)
Quality	-0.5465***	-0.9462***	0.2352***	0.1368
	(0.0744)	(0.0843)	(0.09)	(0.0947)
BIC	0.8897***	0.7699***	-0.3279***	-0.5675***
	(0.0744)	(0.0843)	(0.09)	(0.0947)
Breakpoints	-0.0326	-0.4676***	-0.1934**	0.5415***
	(0.0744)	(0.0843)	(0.09)	(0.0947)
Constant	0.2477*	0.5053***	0.3896**	-0.3620**
	(0.1279)	(0.1447)	(0.1546)	(0.1626)
Observations	352	352	352	352
Adjusted R^2	0.4192	0.4394	0.5391	0.2308

Notes: Standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

In conclusion, the meta-regression results support the findings obtained in the previous sections while controlling for various characteristics of the estimation process. Using quality-adjusted energy use reduces *p*-values substantially for tests of Granger causality from energy use to GDP. This holds also for the use of capital as a control variable, while there is no support for the importance of labor. Note that all of the regressors considered in the meta-regression are theoretically motivated and this analysis should not be misinterpreted as a guide to *p*-hacking (Simonsohn et al., 2014; Bruns and Ioannidis, 2016).

5. Conclusions

Stern (1993) finds Granger causality from energy use to GDP if (i) quality-adjusted energy is used and (ii) capital and labor are included as control variables, whereas Granger causality tests with unadjusted energy use or without capital and labor as control variables are non-significant. We can verify these main inferences using the data compiled for Stern (2000), which is similar to the original data, as the original data are not available anymore. Applying the original estimation approach to revised data published by the relevant US statistical agencies alters the inferences for both the original time span and an extended time span until 2015. However, Granger causality tests based on VARs in levels, as implemented in Stern (1993), are biased, if the underlying time series are non-stationary (Toda and Philips, 1993). Using the Toda and Yamamoto (1995) procedure we find that the original inferences

remain largely stable for the updated data set for both the (almost) original time span (1949-1990) and the extended time span (1949-2015). Specifically, the Granger causality tests from quality-adjusted final energy use to GDP become only marginally non-significant at the 0.1 level while all other inferences remain stable. If structural breaks are considered, the original inferences can be confirmed for both the (almost) original time span and the updated time span for the Toda-Yamamoto procedure and even for the original estimation approach. Overall, these findings tend to support Stern's (1993 two conclusions.

The comprehensive robustness analysis shows that the original inferences are not what we term strongly robust with respect to alternative variable definitions and VAR specifications, that is, Granger causality tests from energy use to GDP need to be exclusively statistically significant if quality-adjusted energy use, capital, and labor are used while all other VAR specifications result in exclusively non-significant *p*-values. However, we do find strong evidence that both quality-adjustment of energy use and the use of capital and to some extent the use of labor as control variables substantially reduces the *p*-values of tests of Granger causality from energy to GDP. This tends to support Stern's (1993) first conclusion that quality-adjusting energy use is essential to finding Granger causality from energy use to GDP. Using joules to measure the energy input results in error in the measurement of the contribution of energy to production as this depends on more than just the heat content of the energy carriers. Measurement error usually biases estimates of regression coefficients towards zero (e.g. Hausman, 2001) that then may result in non-significant Granger causality tests. Our results also partially support the second claim of Stern (1993). We find that capital is a key control variable that reduces the *p*-values of Granger causality tests from energy use to GDP, but the evidence on the role of labor as a control variable is more mixed. These results seem to be intuitive as energy is a production factor and assessing its contribution to economic output requires controlling for the other main production factors. Otherwise, omitted-variable biases are likely to be present.

Comparing the re-analyses of the (almost) original time span (1949-1990) and the extended time span (1949-2015) permits us to speculate about the stability of the data-generating process. If the data-generating process is stable and the VAR well-specified, then the *p*-values should decrease as the sample size increases. We do not find decreasing *p*-values though the variation of *p*-values decreases. If structural breaks are considered, the *p*-values indeed systematically decrease for VARs with quality-adjusted energy use and capital and labor as control variables if the sample size increases. This suggests that there may be structural breaks in the data-generating process which is consistent with previous studies (e.g. Rodríguez-Caballero and Ventosa-Santaulària, 2016).

This replication study includes the original author in designing the replication and robustness analysis using a pre-analysis plan. This is intended to address the incentive for replicating authors to make a study look non-replicable or fragile in order to increase the probability of getting published. While the analysis was conducted independently of the original author and closely following the pre-analysis plan, the final results were interpreted jointly. Given that the scope of a replication is clearly defined by the

original study, it is easy to formulate a pre-analysis plan once agreement is found on which robustness analyses are reasonable. Even disagreements on reasonable robustness checks could be transparently documented in the pre-analysis plan.

Finally, this study is a replication of a study that was published about 25 years ago. This type of replication created major challenges in the verification of the original results, as only revised data are available. However, our results suggest that these data revisions do not alter the inferences substantially.

Our findings lead us to renew the call for better accounting for changes in the energy mix in time series modeling of the energy-GDP relationship. So far, only a few studies use quality-adjusted energy use to assess the role of energy in economic growth (e.g. Oh and Lee, 2004; Shahiduzzaman and Alam, 2010).[10]

References

Bruns S. B., Gross, C., Stern, D. I., 2014. Is there really Granger causality between energy use and output? *Energy Journal* 35(4), 101-134.

Bruns S. B., Ioannidis, J. P. A., 2016. *p*-curve and *p*-hacking in observational research. *PLoS ONE* 11(2), e0149144.

Bruns, S. B., 2017. Meta-Regression models and observational research. *Oxford Bulletin of Economics and Statistics* 79(5), 637-653.

Bruns, S. B., König, J., Stern, D. I., 2017. Pre-analysis plan of replication, update, and robustness analysis of 'Energy and economic growth in the USA: a multivariate approach' (ID: 20170325AA). http://egap.org/registration/2432

Bruns, S. B., Stern, D. I., in press. Lag length selection and *p*-hacking in Granger causality testing: Prevalence and performance of meta-regression models. *Empirical Economics*.

Clarke, J. A., Mirza, S., 2006. A comparison of some common methods for detecting Granger noncausality. *Journal of Statistical Computation and Simulation* 76(3), 207-231.

Clemens, M. A., 2017. The meaning of failed replications: A review and proposal. *Journal of Economic Surveys* 31(1): 326–342.

Cumming, G., 2014. The new statistics: Why and how. *Psychological Science* 25(1), 7-29.

Dewald, W. G., 1986. Replication in empirical economics: The Journal of Money, Credit and Banking Project. *American Economic Review* 78(4), 587-603.

[10] A few studies (e.g. Warr and Ayres, 2009) adjust energy use for useful work, which is an alternative way of adjusting for the productivity of different fuels and electricity.

Duvendack, M., Palmer-Jones, R. W., Reed, W. R., 2015. Replications in economics: A progress report. *Econ Journal Watch* 12(2), 164-191.

Fox. J., Weisberg, S., 2011. An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage

Galiani, S., Gertler, P., Romero, M., 2017. Incentives for replication in economics. *NBER Working Paper 23576.*

Hausman, J., 2001. Mismeasured variables in econometric analysis: problems from the right and problems from the left, *Journal of Economic Perspectives* 54(4), 57-67.

Kilian, L., Lütkepohl, H., 2017. *Structural Vector Autoregressive Analysis*. Cambridge University Press.

Leamer, E. E., 1983. Let's take the con out of econometrics. *The American Economic Review* 73(1), 31-43.

Leamer, E. E., Leonard, H., 1983. Reporting the fragility of regression estimates. *Review of Economics and Statistics* 65(2), 306-317.

Lütkepohl, H., 1982. Non-causality due to omitted variables. *Journal of Econometrics* 19(2-3), 367-378.

Lütkepohl, H., 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Oh, W., Lee, K., 2004. Causal relationship between energy consumption and GDP revisited: the case of Korea 1970– 1999. *Energy Economics* 26, 51–59.

Pfaff, B., 2008. VAR, SVAR and SVEC Models: Implementation Within R Package vars. *Journal of Statistical Software* 27(4).

R Core Team 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodríguez-Caballero, C. V., Ventosa-Santaulària, D, 2016. Energy-growth long-term relationship under structural breaks. Evidence from Canada, 17 Latin American economies and the USA. *Energy Economics* 61:121-134.

Shahiduzzaman, M., Alam, K., 2012. Cointegration and causal relationships between energy consumption and output: Assessing the evidence from Australia. *Energy Economics* 34(6), 2182-2188.

Simonsohn, U., Nelson, L. D., Simmons, J. P., 2014. *p*-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* 143(2): 534-547.

Stern, D. I., 1993. Energy and economic growth in the USA: a multivariate approach. *Energy Economics* 15(2), 137-150.

Stern, D. I., 2000. A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Economics* 22, 267-283.

Stern, D. I., 2010. Energy quality, *Ecological Economics* 69(7), 1471-1478.

Toda, H. Y., Phillips, P. C. B., 1993. The spurious effect of unit roots on vector autoregressions, *Journal of Econometrics* 59(3), 229-255.

Toda, H. Y., Yamamoto, T., 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* 66(1), 225-250.

U.S. Energy Information Administration, Independent Statistics and Analysis, EIA, 2017. Monthly Energy Review – March 2017, Government Printing Office, Washington, D.C.

U.S. Department of Commerce, Bureau of Economic Analysis, BEA, 2017a. National Data – Fixed Assets, https://www.bea.gov/iTable/iTable.cfm?ReqID=10&step=1#reqid=10&step=1&isuri=1&1003=16&1004=1925&1005=2015 (Downloaded, April 2017).

U.S. Department of Commerce, Bureau of Economic Analysis, BEA, 2017b. National Data – GDP & Personal Income, https://bea.gov/iTable/iTable.cfm?ReqID=9#reqid=9&step=1&isuri=1 (Downloaded, April 2017).

U.S. Department of Commerce, Bureau of Economic Analysis, BEA, 2017c. National Economic Accounts - Gross Domestic Product (GDP), https://www.bea.gov/national/index.htm#gdp (Downloaded, April 2017).

U.S. Department of Labor, Bureau of Labour Statistic, DOL, 2017. Labor Force Statistics from the Current Population Survey, https://data.bls.gov/pdq/SurveyOutputServlet (Downloaded, April 2017).

Warr, B. S., Ayres, R. U., 2009. Evidence of causality between the quantity and quality of energy consumption and economic growth. *Energy* 35(4), 1688-1693.

Appendix A

Table A1: Data used in the reanalysis and extension

Where not otherwise stated the data source is EIA (2017)

GDP	Gross domestic product in 2009 $ using price indexes for gross domestic product. Data source: BEA (2017c).
Capital without residential	Private and government capital in 2009 $ using chain-type quantity indexes for net stock of fixed assets and consumer durable goods. Data source: BEA (2017a).
Capital with residential	Private, government and residential capital in 2009 $ using chain-type quantity indexes for net stock of fixed assets and consumer durable goods. Data source: BEA (2017a).
Capital without residential (adjusted by utilization rate)	Capital without residential capital, as outlined above, adjusted by the utilization rate. Utilization rate is defined as (1-unemploymen rate). Data source for unemployment rate: DOL (2017).
Capital with residential (adjusted by utilization rate)	Capital with residential capital, as outlined above, adjusted by the utilization rate. Utilization rate is defined as (1-unemploymen rate). Data source for unemployment rate: DOL (2017).
FTE employment	Full-time equivalent employment. Data source: BEA (2017b).
Hours worked	Hours worked. Data source: BEA (2017b).
Primary energy use	Primary energy use for the following sources: petroleum, natural gas, coal, hydroelectric power, nuclear electric power, geothermal, biomass, wood energy, waste energy, biofuels consumption, solar, and wind energy. Corrected for coal coke exports and net electricity generation from hydroelectric pumped storage. Missing values were extrapolated backwards using growth rates from Stern (2000).
Final energy use	Final energy use constructed by subtracting from primary energy consumption electric power sector energy consumption, adding electricity retail sales, subtracting purchase from nonutility generation (net generation of commercial sector and industrial sector), and subtracting electricity net imports.
Quality-adjusted final energy use	Quality-adjusted final energy use obtained using the Divisia index as described in Stern (1993). Instead of using 60% of the coal price for biomass and the ‘other’ category, we use primary biomass prices.

Quality-adjusted primary energy use	Quality-adjusted primary energy use. Divisia index estimated by method described in Stern (1993). Instead of using 60% of the coal price for biomass and the 'other' category, we use primary biomass prices.
Primary energy prices	Primary energy price of one British thermal unit (BTU) in 2009 $ obtained dividing total expenditures (obtained by multiplying BTUs of all primary energy types with their prices) by total primary BTUs. The prices used were crude oil domestic first purchase price, natural gas price (Wellhead), coal prices (total), and average retail price of electricity (industrial). After 1970, the price of wood and waste was used as biomass price. Before 1970, the price of lumber was used as biomass price. For years after 2012, the wellhead natural gas price is not available. We extrapolate the natural gas price by using the growth rates of Citygate natural gas price. Price information are measured in the *Monthly Energy Review* in $ per Kilowatthour/Short Ton/Barrel/Cubic Feet. To convert them in prices per BTU we used the British Thermal Unit Conversion Factors from the *Monthly Energy Review* (March 2017). These conversion factors are computed on final energy data since no other information are available.
Final energy prices	Final energy price of one British thermal unit (BTU) in 2009 $ obtained by dividing total expenditures (obtained by multiplying BTUs of all final energy types with their prices) by total final BTUs. The prices used were natural gas price (delivered to Consumers Industrial), cost of coal receipts at electric generating plants, average retail price of electricity (total), and the cost of residential heating oil to end users. Before the 1970s, we extrapolated prices backwards using the growth rates of the Stern (2000) time series if the prices were not available in EIA (2017). After 1970, the price of wood and waste was used as the price of biomass. Before 1970, the price of lumber was used as the biomass price. Based on Stern (2000), fossil fuel prices are improved using the expenditure data reported in the US Energy Information Administration's *Annual Energy Review* to obtain better estimates of actual final use fuel prices for oil, natural gas, and coal.
Quality-adjusted primary energy prices	Quality-adjusted simple energy price in 2009 $. Divisia index calculated as described in Stern (1993). Instead of using 60% of the coal price for biomass and the 'other' category, we use primary biomass prices.

Quality adjusted final energy prices	Quality-adjusted final energy price in 2009 $. Divisia index calculated as described in Stern (1993). Instead of using 60% of the coal price for biomass and the 'other' category, we use primary biomass prices.

Appendix B

1. Full results for the multivariate models of Section 2 (Verification)

We use tables in the same style as used in Stern (1993) to ease the comparison.

Table 1.1: Verification of Stern (1993) Table 6 (Primary energy use)

	Dependent Variables			
	GDP	Labor	Capital	Energy
GDP	60.4658	24.0083	26.5603	8.0747
	0	*0.000000*	*0.000000*	*0.0014*
Labor	12.5002	22.6164	15.4970	14.1385
	0.0001	*0.000001*	*0.00002*	*0.00004*
Capital	9.3305	9.5747	26.6177	1.1262
	0.0006	*0.0005*	*0.000000*	*0.3364*
Energy	2.2994	1.0224	1.4838	39.2932
	0.1162	*0.3708*	*0.2415*	*0*
R^2	0.9985	0.9966	0.9994	0.9938

Notes: Stern (2000) data and original timespan. Variables are in log levels and two lags are used. *p*-values in italics. Each test statistic is for a test of Granger causality from the variable in the first column to the dependent variable.

Table 2.1: Verification of Stern (1993) Table 10 (Quality-adjusted final energy use)

	Dependent Variables			
	GDP	Labor	Capital	Energy
GDP	21.5368	5.9653	6.7493	1.0851
	0.000000	*0.0019*	*0.0010*	*0.3870*
Labor	6.6079	1.2207	3.5809	2.2040
	0.0011	*0.3293*	*0.0207*	*0.1003*
Capital	6.2118	4.3427	5.2347	1.4893
	0.0015	*0.0092*	*0.0038*	*0.2381*
Energy	5.0755	2.9697	2.4370	23.1178
	0.0044	*0.0409*	*0.0760*	*0.000000*
R^2	0.9992	0.9976	0.9996	0.9975

Notes: Stern (2000) data and original timespan. Variables are in log levels and four lags are used. 60% of the coal price is used for the prices of biomass and the 'other' category. Variables are in log levels. *p*-values in italics. Each test statistic is for a test of Granger causality from the variable in the first column to the dependent variable.

2. Unit root tests

Table 2.1: Overview of the variable names

GDP	GDP09
Capital without residential	KPG09
Capital with residential	KPGR09
Capital without residential (adjusted by utilization rate)	KPG09UN
Capital with residential (adjusted by utilization rate)	KPGR09UN
FTE employment	FULLEMPNEW
Hours worked	HOURSNEW
Primary energy use	E
Final energy use	FINENEW
Quality-adjusted final energy use	QEFDIVFINNEW
Quality-adjusted primary energy use	QEFDIVPRIMNEW
Primary energy prices	PSIMPPRIMNEW
Final energy prices	PSIMPFINNEW
Quality-adjusted primary energy prices	PDIVFINNEW
Quality adjusted final energy prices	PDIVPRIMNEW

Table 2.2: Augmented Dickey-Fuller Tests, 1949 - 2015 (AIC)

Variable		Mod1		Mod2			Mod3						Mod1	Mod2	Mod3
		τ_1		τ_2		ϕ_1	τ_3		ϕ_2		ϕ_3		Lags	Lags	Lags
Δ(log(GDP09))	AIC	-1.6846	*	-5.0814	***	12.9226	***	-5.4212	***	9.8104	***	14.7028	***	3	1	1
Δ(log(KPG09UN))	AIC	-1.287		-4.4909	***	10.1033	***	-5.2288	***	9.1512	***	13.7061	***	3	1	1
Δ(log(KPGR09UN))	AIC	-1.4019		-4.5985	***	10.5908	***	-5.4513	***	9.9351	***	14.8831	***	3	1	1
Δ(log(KPG09))	AIC	-1.3543		-1.6483		1.7985		-4.2065	***	6.1106	**	8.8558	***	2	2	1
Δ(log(KPGR09))	AIC	-1.3537		-1.0565		1.1639		-4.001	**	5.5664	**	8.0048	**	3	3	1
Δ(log(FULLEMPNEW))	AIC	-2.3149	**	-5.6955	***	16.2195	***	-5.8923	***	11.5816	***	17.3722	***	3	1	1
Δ(log(HOURSNEW))	AIC	-4.4553	***	-6.1841	***	19.1223	***	-6.1886	***	12.7711	***	19.1559	***	1	1	1
Δ(log(E))	AIC	-2.1458	**	-4.7329	***	11.2086	***	-5.7825	***	11.1682	***	16.743	***	3	1	1
Δ(log(FINE))	AIC	-2.5034	**	-5.1096	***	13.0563	***	-5.8309	***	11.3414	***	17.0098	***	3	1	1
Δ(log(QEFDIVFIN))	AIC	-2.0931	**	-2.9329	**	4.3457	*	-5.5829	***	10.4028	***	15.5864	***	3	2	1
Δ(log(QEFDIVPRIM))	AIC	-2.5152	**	-4.2611	***	9.0788	***	-5.0278	***	8.4305	***	12.6454	***	2	1	1
Δ(log(PSIMPPRIM))	AIC	-3.2777	***	-3.2463	**	5.3436	**	-3.1364		3.5539		5.2574		2	2	2
Δ(log(PSIMPFIN))	AIC	-4.6207	***	-4.6757	***	11.0238	***	-4.6128	***	7.2585	***	10.7963	***	1	1	1
Δ(log(PDIVFIN))	AIC	-4.8891	***	-4.8627	***	11.8983	***	-4.7862	***	7.8115	***	11.6432	***	1	1	1
Δ(log(PDIVPRIM))	AIC	-3.3183	***	-3.2865	**	5.481	**	-3.177	*	3.6379		5.3776		2	2	2
log(GDP09)	AIC	5.3138		-1.8232		16.8184	***	-0.9023		11.309	***	1.922		1	1	1
log(KPG09UN)	AIC	4.3977		-2.4034		13.7927	***	-0.9377		9.244	***	3.0677		1	1	1
log(KPGR09UN)	AIC	4.3351		-2.4679		13.6707	***	-0.9133		9.1438	***	3.2011		1	1	1
log(KPG09)	AIC	1.8788		-3.3959	**	8.1044	***	0.0964		5.5261	**	5.9704	*	2	2	2
log(KPGR09)	AIC	1.5574		-3.3474	**	7.208	***	0.1653		4.9958	**	5.8976	*	2	2	2
log(FULLEMPNEW)	AIC	3.8881		-1.5744		9.1666	***	-0.6873		6.0897	**	1.3164		2	2	2
log(HOURSNEW)	AIC	3.7094		-0.9888		7.4248	***	-1.4335		5.5644	**	1.3313		2	2	2
log(E)	AIC	2.9641		-3.0318	**	9.675	***	-1.0476		6.3451	**	4.5229		1	1	1
log(FINE)	AIC	2.5603		-2.7976	*	7.6129	***	-1.5832		5.0613	**	3.9436		1	1	1
log(QEFDIVFIN)	AIC	1.7045		-3.5756	***	10.6056	***	-1.5637		6.9541	***	6.288	*	3	1	1
log(QEFDIVPRIM)	AIC	2.5143		-2.5711		6.824	***	-1.0599		4.4774	*	3.2554		1	1	1
log(PSIMPPRIM)	AIC	-0.523		-1.5416		1.2991		-2.5458		2.2507		3.3147		1	1	3
log(PSIMPFIN)	AIC	-0.9954		-1.3597		1.3853		-2.1222		1.8676		2.327		1	1	1
log(PDIVFIN)	AIC	-0.5163		-1.4818		1.2177		-2.2552		1.7791		2.5451		1	1	1
log(PDIVPRIM)	AIC	-0.5417		-1.56		1.3369		-2.5732		2.2991		3.3798		1	1	3

Augmented Dickey-Fuller unit root test. Null hypothesis is the presence of a unit root.
* p<0.1; ** p<0.05; *** p<0.01
Model 1: Neither an intercept nor a trend is included in the test regression.
Model 2: An intercept is included in the test regression.
Model 3: Intercept and a trend is included in the test regression
ADF-test equation: $\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma y_{t-1} + \mu_t$

$\tau_1 : \pi = 0$ $\quad$ $\phi_1 : \beta_1 = \pi = 0$
$\tau_2 : \pi = 0$ $\quad$ $\phi_2 : \beta_1 = \pi = \beta_2 = 0$
$\tau_3 : \pi = 0$ $\quad$ $\phi_3 : \beta_2 = \pi = 0$

Table 2.3: Augmented Dickey-Fuller Tests, 1949 - 2015 (BIC)

Variable		Mod1		Mod2			Mod3					Mod1	Mod2	Mod3
		τ_1		τ_2		ϕ_1	τ_3		ϕ_2		ϕ_3	Lags	Lags	Lags
Δ(log(GDP09))	BIC	-2.244	**	-5.0814	***	12.9226 ***	-5.4212	***	9.8104	***	14.7028 ***	1	1	1
Δ(log(KPG09UN))	BIC	-1.287		-4.4909	***	10.1033 ***	-5.2288	***	9.1512	***	13.7061 ***	3	1	1
Δ(log(KPGR09UN))	BIC	-1.9634	**	-4.5985	***	10.5908 ***	-5.4513	***	9.9351	***	14.8831 ***	1	1	1
Δ(log(KPG09))	BIC	-1.3543		-1.6483		1.7985	-4.2065	***	6.1106	**	8.8558 ***	2	2	1
Δ(log(KPGR09))	BIC	-1.3331		-2.0608		2.418	-4.001	**	5.5664	**	8.0048 **	1	1	1
Δ(log(FULLEMPNEW))	BIC	-3.6266	***	-5.6955	***	16.2195 ***	-5.8923	***	11.5816	***	17.3722 ***	1	1	1
Δ(log(HOURSNEW))	BIC	-4.4553	***	-6.1841	***	19.1223 ***	-6.1886	***	12.7711	***	19.1559 ***	1	1	1
Δ(log(E))	BIC	-2.6469	***	-4.7329	***	11.2086 ***	-5.7825	***	11.1682	***	16.743 ***	2	1	1
Δ(log(FINE))	BIC	-4.308	***	-5.1096	***	13.0563 ***	-5.8309	***	11.3414	***	17.0098 ***	1	1	1
Δ(log(QEFDIVFIN))	BIC	-2.34	**	-4.1537	***	8.642 ***	-5.5829	***	10.4028	***	15.5864 ***	2	1	1
Δ(log(QEFDIVPRIM))	BIC	-3.4312	***	-4.2611	***	9.0788 ***	-5.0278	***	8.4305	***	12.6454 ***	1	1	1
Δ(log(PSIMPPRIM))	BIC	-5.0815	***	-5.0603	***	12.8809 ***	-4.9727	***	8.463	***	12.6181 ***	1	1	1
Δ(log(PSIMPFIN))	BIC	-4.6207	***	-4.6757	***	11.0238 ***	-4.6128	***	7.2585	***	10.7963 ***	1	1	1
Δ(log(PDIVFIN))	BIC	-4.8891	***	-4.8627	***	11.8983 ***	-4.7862	***	7.8115	***	11.6432 ***	1	1	1
Δ(log(PDIVPRIM))	BIC	-5.1667	***	-5.1468	***	13.3273 ***	-5.0577	***	8.7523	***	13.0474 ***	1	1	1
log(GDP09)	BIC	5.3138		-1.8232		16.8184 ***	-0.9023		11.309	***	1.922	1	1	1
log(KPG09UN)	BIC	4.3977		-2.4034		13.7927 ***	-0.9377		9.244	***	3.0677	1	1	1
log(KPGR09UN)	BIC	4.3351		-2.4679		13.6707 ***	-0.9133		9.1438	***	3.2011	1	1	1
log(KPG09)	BIC	1.8788		-3.3959	**	8.1044 ***	0.0964		5.5261	**	5.9704 *	2	2	2
log(KPGR09)	BIC	1.5574		-3.3474	**	7.208 ***	0.1653		4.9958	**	5.8976 *	2	2	2
log(FULLEMPNEW)	BIC	3.4745		-1.2308		6.9571 ***	-1.2624		5.0272	**	1.3272	1	1	1
log(HOURSNEW)	BIC	3.1482		-0.9888		7.4248 ***	-2.0772		4.9677	**	2.2561	1	2	1
log(E)	BIC	2.9641		-3.0318	**	9.675 ***	-1.0476		6.3451	**	4.5229	1	1	1
log(FINE)	BIC	2.5603		-2.7976	*	7.6129 ***	-1.5832		5.0613	**	3.9436	1	1	1
log(QEFDIVFIN)	BIC	2.6215		-3.5756	***	10.6056 ***	-1.5637		6.9541	***	6.288 *	1	1	1
log(QEFDIVPRIM)	BIC	2.5143		-2.5711		6.824 ***	-1.0599		4.4774	*	3.2554	1	1	1
log(PSIMPPRIM)	BIC	-0.523		-1.5416		1.2991	-2.1281		1.6089		2.3003	1	1	1
log(PSIMPFIN)	BIC	-0.9954		-1.3597		1.3853	-2.1222		1.8676		2.327	1	1	1
log(PDIVFIN)	BIC	-0.5163		-1.4818		1.2177	-2.2552		1.7791		2.5451	1	1	1
log(PDIVPRIM)	BIC	-0.5417		-1.56		1.3369	-2.1575		1.6565		2.3621	1	1	1

Augmented Dickey-Fuller unit root test. Null hypothesis is the presence of a unit root.
* p<0.1; ** p<0.05; *** p<0.01
Model 1: Neither an intercept nor a trend is included in the test regression.
Model 2: An intercept is included in the test regression.
Model 3: Intercept and a trend is included in the test regression
ADF-test equation: $\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma y_{t-1} + \mu_t$
$\tau_1 : \pi = 0$ $\phi_1 : \beta_1 = \pi = 0$
$\tau_2 : \pi = 0$ $\phi_2 : \beta_1 = \pi = \beta_2 = 0$
$\tau_3 : \pi = 0$ $\phi_3 : \beta_2 = \pi = 0$

Table 2.4: Augmented Dickey-Fuller Tests, 1949 - 1990 (AIC)

Variable		Mod1		Mod2			Mod3						Mod1	Mod2	Mod3
		τ_1		τ_2		ϕ_1	τ_3		ϕ_2		ϕ_3		Lags	Lags	Lags
Δ(log(GDP09))	AIC	-1.3317		-4.5993	***	10.5952	***	-4.5366	***	6.8721	**	10.2903	***	3	1	1
Δ(log(KPG09UN))	AIC	-0.9487		-4.1534	***	8.6681	***	-4.2198	***	5.9642	**	8.9038	**	3	1	1
Δ(log(KPGR09UN))	AIC	-1.0324		-4.5093	***	10.208	***	-4.6573	***	7.2579	***	10.8455	***	3	1	1
Δ(log(KPG09))	AIC	-1.0376		-1.6911		1.7332		-3.1468		3.4532		4.9567		2	2	1
Δ(log(KPGR09))	AIC	-1.0512		-2.2219		2.7304		-3.674	**	4.7111	*	6.7537	**	3	1	1
Δ(log(FULLEMPNEW))	AIC	-1.636	*	-5.2029	***	13.5375	***	-5.2595	***	9.2229	***	13.8318	***	3	1	1
Δ(log(HOURSNEW))	AIC	-2.0147	**	-5.4129	***	14.6503	***	-5.5774	***	10.3709	***	15.5561	***	3	1	1
Δ(log(E))	AIC	-1.5124		-3.7257	***	6.9448	**	-4.0697	**	5.5537	**	8.3261	**	3	1	1
Δ(log(FINE))	AIC	-3.0241	***	-3.7936	***	7.204	***	-4.2005	***	5.9325	**	8.89	**	1	1	1
Δ(log(QEFDIVFIN))	AIC	-1.5378		-3.528	**	6.2513	**	-4.3772	***	6.4215	**	9.6006	***	3	1	1
Δ(log(QEFDIVPRIM))	AIC	-2.4357	**	-3.4325	**	5.8914	**	-3.6252	**	4.4093	*	6.6134	*	1	1	1
Δ(log(PSIMPPRIM))	AIC	-3.072	***	-3.0708	**	4.7259	*	-3.0241		3.072		4.5974		1	1	1
Δ(log(PSIMPFIN))	AIC	-3.1262	***	-3.1943	**	5.1017	**	-3.1367		3.3108		4.9662		1	1	1
Δ(log(PDIVFIN))	AIC	-3.3031	***	-3.2679	**	5.3401	**	-3.2326	*	3.5012		5.2511		1	1	1
Δ(log(PDIVPRIM))	AIC	-3.1923	***	-3.1874	**	5.0875	**	-3.1365		3.2998		4.9421		1	1	1
log(GDP09)	AIC	4.8608		-0.4617		11.7593	***	-2.0858		9.9606	***	2.2202		1	1	1
log(KPG09UN)	AIC	4.51		-1.0554		10.9851	***	-1.4728		8.1492	***	1.5251		1	1	1
log(KPGR09UN)	AIC	4.6305		-1.2089		11.8417	***	-1.4952		8.749	***	1.7165		1	1	1
log(KPG09)	AIC	2.1373		-2.0837		4.9596	**	-0.4805		3.2495		2.1599		2	2	2
log(KPGR09)	AIC	1.8371		-2.9393	**	6.6908	**	0.0065		4.3726	*	4.2511		2	2	2
log(FULLEMPNEW)	AIC	4.0163		0.3191		7.8774	***	-3.4154	*	9.5432	***	6.0798	*	2	2	1
log(HOURSNEW)	AIC	3.6349		0.5511		6.6018	**	-3.5481	**	8.6827	***	6.829	**	2	2	1
log(E)	AIC	2.7659		-1.761		5.6521	**	-0.9929		3.7539		1.6216		1	1	1
log(FINE)	AIC	2.2618		-1.9174		4.6203	*	-1.0299		3.0278		1.8323		1	1	1
log(QEFDIVFIN)	AIC	2.4988		-2.3467		6.346	**	-1.0081		4.1436		2.7178		1	1	1
log(QEFDIVPRIM)	AIC	2.3922		-1.4563		4.0485	*	-1.2167		2.8922		1.3794		1	1	1
log(PSIMPPRIM)	AIC	-0.5667		-1.228		0.8949		-1.982		1.4149		1.9758		1	1	1
log(PSIMPFIN)	AIC	-0.8967		-1.1183		1.0049		-2.0002		1.599		2.0005		1	1	1
log(PDIVFIN)	AIC	-0.3755		-1.4394		1.0992		-1.9787		1.3579		1.9721		1	1	1
log(PDIVPRIM)	AIC	-0.5642		-1.2718		0.9489		-1.9912		1.4209		1.9861		1	1	1

Augmented Dickey-Fuller unit root test. Null hypothesis is the presence of a unit root.
* p<0.1; ** p<0.05; *** p<0.01
Model 1: Neither an intercept nor a trend is included in the test regression.
Model 2: An intercept is included in the test regression.
Model 3: Intercept and a trend is included in the test regression
ADF-test equation: $\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma y_{t-1} + \mu_t$
$\tau_1 : \pi = 0$ $\phi_1 : \beta_1 = \pi = 0$
$\tau_2 : \pi = 0$ $\phi_2 : \beta_1 = \pi = \beta_2 = 0$
$\tau_3 : \pi = 0$ $\phi_3 : \beta_2 = \pi = 0$

Table 2.5: Augmented Dickey-Fuller Tests, 1949 - 1990 (BIC)

Variable		Mod1		Mod2			Mod3					Mod1	Mod2	Mod3
		τ_1		τ_2		ϕ_1	τ_3		ϕ_2		ϕ_3		Lags	Lags	Lags
Δ(log(GDP09))	BIC	-1.8121	*	-4.5993	***	10.5952	***	-4.5366	***	6.8721	**	10.2903	***	1	1	1
Δ(log(KPG09UN))	BIC	-0.9487		-4.1534	***	8.6681	***	-4.2198	***	5.9642	**	8.9038	**	3	1	1
Δ(log(KPGR09UN))	BIC	-1.0324		-4.5093	***	10.208	***	-4.6573	***	7.2579	***	10.8455	***	3	1	1
Δ(log(KPG09))	BIC	-1.0376		-2.4908		3.311		-3.1468		3.4532		4.9567		2	1	1
Δ(log(KPGR09))	BIC	-1.0478		-2.2219		2.7304		-3.674	**	4.7111	*	6.7537	**	1	1	1
Δ(log(FULLEMPNEW))	BIC	-2.6845	***	-5.2029	***	13.5375	***	-5.2595	***	9.2229	***	13.8318	***	1	1	1
Δ(log(HOURSNEW))	BIC	-3.4269	***	-5.4129	***	14.6503	***	-5.5774	***	10.3709	***	15.5561	***	1	1	1
Δ(log(E))	BIC	-2.6082	**	-3.7257	***	6.9448	**	-4.0697	**	5.5537	**	8.3261	**	1	1	1
Δ(log(FINE))	BIC	-3.0241	***	-3.7936	***	7.204	***	-4.2005	***	5.9325	**	8.89	**	1	1	1
Δ(log(QEFDIVFIN))	BIC	-2.3762	**	-3.528	**	6.2513	**	-4.3772	***	6.4215	**	9.6006	***	1	1	1
Δ(log(QEFDIVPRIM))	BIC	-2.4357	**	-3.4325	**	5.8914	**	-3.6252	**	4.4093	*	6.6134	*	1	1	1
Δ(log(PSIMPPRIM))	BIC	-3.072	***	-3.0708	**	4.7259	*	-3.0241		3.072		4.5974		1	1	1
Δ(log(PSIMPFIN))	BIC	-3.1262	***	-3.1943	**	5.1017	**	-3.1367		3.3108		4.9662		1	1	1
Δ(log(PDIVFIN))	BIC	-3.3031	***	-3.2679	**	5.3401	**	-3.2326	*	3.5012		5.2511		1	1	1
Δ(log(PDIVPRIM))	BIC	-3.1923	***	-3.1874	**	5.0875	**	-3.1365		3.2998		4.9421		1	1	1
log(GDP09)	BIC	4.8608		-0.4617		11.7593	***	-2.0858		9.9606	***	2.2202		1	1	1
log(KPG09UN)	BIC	4.51		-1.0554		10.9851	***	-1.4728		8.1492	***	1.5251		1	1	1
log(KPGR09UN)	BIC	4.6305		-1.2089		11.8417	***	-1.4952		8.749	***	1.7165		1	1	1
log(KPG09)	BIC	1.5688		-2.0837		4.9596	**	-0.4805		3.2495		2.1599		1	2	2
log(KPGR09)	BIC	1.4118		-2.9393	**	6.6908	**	0.0065		4.3726	*	4.2511		1	2	2
log(FULLEMPNEW)	BIC	3.5896		0.2732		6.2948	**	-3.4154	*	9.5432	***	6.0798	*	1	1	1
log(HOURSNEW)	BIC	3.0636		0.4162		4.6588	*	-3.5481	**	8.6827	***	6.829	**	1	1	1
log(E)	BIC	2.7659		-1.761		5.6521	**	-0.9929		3.7539		1.6216		1	1	1
log(FINE)	BIC	2.2618		-1.9174		4.6203	*	-1.0299		3.0278		1.8323		1	1	1
log(QEFDIVFIN)	BIC	2.4988		-2.3467		6.346	**	-1.0081		4.1436		2.7178		1	1	1
log(QEFDIVPRIM)	BIC	2.3922		-1.4563		4.0485	*	-1.2167		2.8922		1.3794		1	1	1
log(PSIMPPRIM)	BIC	-0.5667		-1.228		0.8949		-1.982		1.4149		1.9758		1	1	1
log(PSIMPFIN)	BIC	-0.8967		-1.1183		1.0049		-2.0002		1.599		2.0005		1	1	1
log(PDIVFIN)	BIC	-0.3755		-1.4394		1.0992		-1.9787		1.3579		1.9721		1	1	1
log(PDIVPRIM)	BIC	-0.5642		-1.2718		0.9489		-1.9912		1.4209		1.9861		1	1	1

Augmented Dickey-Fuller unit root test. Null hypothesis is the presence of a unit root.
* p<0.1; ** p<0.05; *** p<0.01
Model 1: Neither an intercept nor a trend is included in the test regression.
Model 2: An intercept is included in the test regression.
Model 3: Intercept and a trend is included in the test regression
ADF-test equation: $\Delta y_t = \beta_1 + \beta_2 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma y_{t-1} + \mu_t$
$\tau_1 : \pi = 0$ $\phi_1 : \beta_1 = \pi = 0$
$\tau_2 : \pi = 0$ $\phi_2 : \beta_1 = \pi = \beta_2 = 0$
$\tau_3 : \pi = 0$ $\phi_3 : \beta_2 = \pi = 0$

Table 2.6: Kwiatkowski-Phillips-Schmidt-Shin Tests, 1949 - 1990

Variable	Lags	Mod1 ημ		Mod2 ητ	
Δ(log(GDP09))	3	0.2015		0.0563	
Δ(log(KPG09UN))	3	0.2422		0.0656	
Δ(log(KPGR09UN))	3	0.3536	*	0.0583	
Δ(log(KPG09))	3	0.4382	*	0.1177	
Δ(log(KPGR09))	3	0.7579	* * *	0.1126	
Δ(log(FULLEMPNEW))	3	0.0602		0.0521	
Δ(log(HOURSNEW))	3	0.0591		0.0567	
Δ(log(E))	3	0.4251	*	0.0801	
Δ(log(FINE))	3	0.4644	* *	0.0722	
Δ(log(QEFDIVFIN))	3	0.6614	* *	0.066	
Δ(log(QEFDIVPRIM))	3	0.2817		0.0892	
Δ(log(PSIMPPRIM))	3	0.118		0.103	
Δ(log(PSIMPFIN))	3	0.0993		0.0992	
Δ(log(PDIVFIN))	3	0.1192		0.0988	
Δ(log(PDIVPRIM))	3	0.1065		0.0978	
log(GDP09)	3	1.154	* * *	0.1851	* *
log(KPG09UN)	3	1.1492	* * *	0.2277	* * *
log(KPGR09UN)	3	1.1501	* * *	0.253	* * *
log(KPG09)	3	1.1489	* * *	0.2622	* * *
log(KPGR09)	3	1.1496	* * *	0.2835	* * *
log(FULLEMPNEW)	3	1.1582	* * *	0.0705	
log(HOURSNEW)	3	1.1568	* * *	0.0644	
log(E)	3	1.0938	* * *	0.2491	* * *
log(FINE)	3	1.0499	* * *	0.2583	* * *
log(QEFDIVFIN)	3	1.0932	* * *	0.2725	* * *
log(QEFDIVPRIM)	3	1.1004	* * *	0.2153	* *
log(PSIMPPRIM)	3	0.68	* *	0.1588	* *
log(PSIMPFIN)	3	0.8589	* * *	0.1496	* *
log(PDIVFIN)	3	0.5211	* *	0.1735	* *
log(PDIVPRIM)	3	0.6915	* *	0.1527	* *

KPSS unit root test. Null hypothesis is stationarity.

* p<0.1; ** p<0.05; *** p<0.01

ημ is the test statistic against levels stationarity, ητ is the test statistic against trend stationarity.

Table 2.7: Kwiatkowski-Phillips-Schmidt-Shin Tests, 1949 - 2015

Variable	Lags	Mod1 ημ		Mod2 ητ	
Δ(log(GDP09))	3	0.543	* *	0.0443	
Δ(log(KPG09UN))	3	0.7117	* *	0.0368	
Δ(log(KPGR09UN))	3	0.7858	* * *	0.0356	
Δ(log(KPG09))	3	1.1633	* * *	0.0783	
Δ(log(KPGR09))	3	1.3017	* * *	0.0759	
Δ(log(FULLEMPNEW))	3	0.4068	*	0.0434	
Δ(log(HOURSNEW))	3	0.2237		0.0383	
Δ(log(E))	3	0.846	* * *	0.06	
Δ(log(FINE))	3	0.6993	* *	0.0763	
Δ(log(QEFDIVFIN))	3	1.0621	* * *	0.0883	
Δ(log(QEFDIVPRIM))	3	0.6868	* *	0.0536	
Δ(log(PSIMPPRIM))	3	0.0728		0.0731	
Δ(log(PSIMPFIN))	3	0.066		0.0639	
Δ(log(PDIVFIN))	3	0.0763		0.0642	
Δ(log(PDIVPRIM))	3	0.0691		0.0694	
log(GDP09)	3	1.7659	* * *	0.3214	* * *
log(KPG09UN)	3	1.7569	* * *	0.3863	* * *
log(KPGR09UN)	3	1.7562	* * *	0.3841	* * *
log(KPG09)	3	1.7535	* * *	0.4245	* * *
log(KPGR09)	3	1.7527	* * *	0.4215	* * *
log(FULLEMPNEW)	3	1.7553	* * *	0.3467	* * *
log(HOURSNEW)	3	1.7629	* * *	0.2692	* * *
log(E)	3	1.5965	* * *	0.3829	* * *
log(FINE)	3	1.5215	* * *	0.3621	* * *
log(QEFDIVFIN)	3	1.5718	* * *	0.3968	* * *
log(QEFDIVPRIM)	3	1.5959	* * *	0.3783	* * *
log(PSIMPPRIM)	3	1.0977	* * *	0.0914	
log(PSIMPFIN)	3	1.3954	* * *	0.0885	
log(PDIVFIN)	3	0.9921	* * *	0.1129	
log(PDIVPRIM)	3	1.0975	* * *	0.0918	

KPSS unit root test, where the Null hypothesis is stationarity.

* p<0.1; ** p<0.05; *** p<0.01

ημ is the test statistic against levels stationarity, ητ is the test statistic against trend stationarity.

3. Diagnostic tests for Table 1 and 2

Table 3.1: Diagnostic tests for multivariate models with capital and labor

		Original estimation approach					Toda-Yamamoto procedure				
Source or data		Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks
		II	III	IV	V	VI	VII	VIII	IX	X	XI
Panel A: Primary energy use											
Energy causes GDP	Jarque-Bera test	0.433 (0.8053)	1.2723 (0.5293)	0.668 (0.7161)	8.8462 (0.012)	2.4405 (0.2952)	1.1272 (0.5692)	0.0779 (0.9618)	0.5755 (0.75)	5.0863 (0.0786)	1.6995 (0.4275)
	OLS-CUSUM	0.4771 (0.9767)	0.4265 (0.9933)	0.3999 (0.9972)	0.4637 (0.9826)	0.4397 (0.9904)	0.4788 (0.9759)	0.2773 (1)	0.3629 (0.9994)	0.3773 (0.9989)	0.4285 (0.9929)
GDP causes energy	Jarque-Bera test	0.4937 (0.7813)	1.1625 (0.5592)	0.2397 (0.8871)	0.2102 (0.9002)	2.1344 (0.344)	1.1886 (0.5519)	1.2651 (0.5312)	1.1561 (0.561)	0.3485 (0.8401)	0.5313 (0.7667)
	OLS-CUSUM	0.5816 (0.8876)	0.5188 (0.9506)	0.4351 (0.9915)	0.4514 (0.987)	0.4713 (0.9794)	0.4372 (0.991)	0.3678 (0.9993)	0.2852 (1)	0.3543 (0.9996)	0.3806 (0.9987)
	Portmanteau test	229.0968 (0.6826)	232.7179 (0.6199)	225.0047 (0.7481)	269.9631 (0.0894)	256.4966 (0.2216)	200.4953 (0.633)	217.3279 (0.3145)	225.1173 (0.1977)	206.1944 (0.5224)	215.1971 (0.3514)
Panel B: Quality-adjusted final energy use											
Energy causes GDP	Jarque-Bera test	0.1952 (0.907)	0.4824 (0.7857)	0.19 (0.9094)	3.6181 (0.1638)	0.19 (0.9094)	0.5572 (0.7569)	0.1952 (0.907)	0.4824 (0.7857)	0.19 (0.9094)	3.6181 (0.1638)
	OLS-CUSUM	0.3771 (0.9989)	0.361 (0.9995)	0.3779 (0.9988)	0.3616 (0.9994)	0.3779 (0.9988)	0.4176 (0.9949)	0.3771 (0.9989)	0.361 (0.9995)	0.3779 (0.9988)	0.3616 (0.9994)
GDP causes energy	Jarque-Bera test	0.6892 (0.7085)	0.39 (0.8228)	0.0369 (0.9817)	1.0063 (0.6046)	0.0369 (0.9817)	0.2167 (0.8973)	0.6892 (0.7085)	0.39 (0.8228)	0.0369 (0.9817)	1.0063 (0.6046)
	OLS-CUSUM	0.5145 (0.9539)	0.4113 (0.9959)	0.3346 (0.9999)	0.347 (0.9997)	0.3346 (0.9999)	0.3519 (0.9997)	0.5145 (0.9539)	0.4113 (0.9959)	0.3346 (0.9999)	0.347 (0.9997)
	Portmanteau test	205.347 (0.539)	214.5805 (0.3625)	221.1335 (0.2535)	205.5919 (0.5342)	221.1335 (0.2535)	222.4934 (0.2335)	205.347 (0.539)	214.5805 (0.3625)	221.1335 (0.2535)	205.5919 (0.5342)

Notes: For the Jarque-Bera tests and Portmanteau tests, χ^2-test statistics and p-values (in parentheses) are reported. For OLS-CUSUM, F-test statistics and p-values (in parentheses) are reported.

Table 3.2: Diagnostic tests for bivariate models

Source or data		Original estimation approach					Toda-Yamamoto procedure				
		Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks	Stern (2000) (1947-1990)	Updated data (1949-1990)	Updated data (1949-1990) with structural breaks	Updated data (1949-2015)	Updated data (1949-2015) with structural breaks
		II	III	IV	V	VI	VII	VIII	IX	X	XI
Panel A: Primary energy use											
Energy causes GDP	Jarque-Bera test	1.801 (0.4064)	2.4241 (0.2976)	2.3091 (0.3152)	5.4629 (0.0651)	4.2978 (0.1166)	1.801 (0.4064)	2.4241 (0.2976)	2.3091 (0.3152)	5.4629 (0.0651)	4.2978 (0.1166)
	OLS-CUSUM	0.4866 (0.9719)	0.5129 (0.9551)	0.4206 (0.9944)	0.4873 (0.9715)	0.4859 (0.9722)	0.4866 (0.9719)	0.5129 (0.9551)	0.4206 (0.9944)	0.4873 (0.9715)	0.4859 (0.9722)
GDP causes energy	Jarque-Bera test	2.7766 (0.2495)	0.6229 (0.7324)	0.0336 (0.9833)	1.5916 (0.4512)	0.2869 (0.8664)	2.7766 (0.2495)	0.6229 (0.7324)	0.0336 (0.9833)	1.5916 (0.4512)	0.2869 (0.8664)
	OLS-CUSUM	0.6644 (0.7695)	0.5487 (0.9241)	0.2696 (1)	0.5024 (0.9624)	0.4884 (0.9709)	0.6644 (0.7695)	0.5487 (0.9241)	0.2696 (1)	0.5024 (0.9624)	0.4884 (0.9709)
	Portmanteau test	43.5366 (0.9458)	35.6666 (0.9947)	42.9405 (0.9528)	41.8093 (0.9644)	44.8953 (0.9269)	43.5366 (0.9458)	35.6666 (0.9947)	42.9405 (0.9528)	41.8093 (0.9644)	44.8953 (0.9269)
Panel B: Quality-adjusted final energy use											
Energy causes GDP	Jarque-Bera test	3.7398 (0.1541)	2.0165 (0.3649)	1.5233 (0.4669)	3.738 (0.1543)	3.0155 (0.2214)	3.7398 (0.1541)	1.9986 (0.3681)	1.9778 (0.372)	3.2729 (0.1947)	2.633 (0.2681)
	OLS-CUSUM	0.5118 (0.9559)	0.4624 (0.9831)	0.3157 (1)	0.4198 (0.9946)	0.3858 (0.9984)	0.5118 (0.9559)	0.519 (0.9505)	0.496 (0.9665)	0.5282 (0.943)	0.4985 (0.9649)
GDP causes energy	Jarque-Bera test	0.8912 (0.6404)	1.4955 (0.4734)	0.5349 (0.7653)	1.9814 (0.3713)	0.7947 (0.6721)	0.8912 (0.6404)	0.4861 (0.7842)	0.7069 (0.7023)	0.8874 (0.6417)	0.1897 (0.9095)
	OLS-CUSUM	0.5177 (0.9515)	0.4983 (0.965)	0.3168 (1)	0.3931 (0.9978)	0.5118 (0.9559)	0.5177 (0.9515)	0.5059 (0.9601)	0.2951 (1)	0.4289 (0.9929)	0.4464 (0.9885)
	Portmanteau test	40.9949 (0.8643)	40.6634 (0.8724)	47.0141 (0.6699)	44.7948 (0.7504)	47.1319 (0.6654)	40.9949 (0.8643)	35.3079 (0.9954)	42.4027 (0.9586)	50.4954 (0.8041)	53.2774 (0.7179)

Notes: For the Jarque-Bera tests and Portmanteau tests, χ^2-test statistics and *p*-values (in parentheses) are reported. For OLS-CUSUM, F-test statistics and *p*-values (in parentheses) are reported.

4. Robustness analysis in first differences

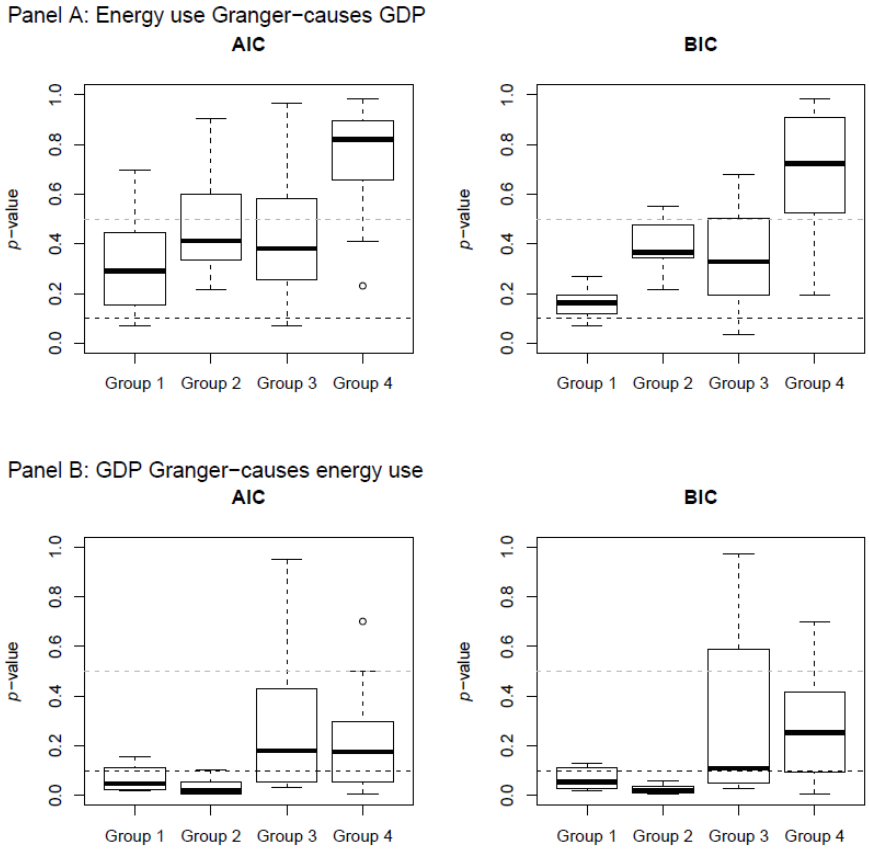

Figure 4.1. Reanalysis of the (almost) original time span 1949-1990 using first differences. Boxplots for *p*-values of Granger causality tests are shown separately for VARs specified with AIC and BIC. Group 1 comprises VARs with quality-adjusted energy use and capital and labor. Group 2 is the same as Group 1 but uses unadjusted energy use. Group 3 comprises VARs with quality adjusted energy use but capital and labor are never included simultaneously as control variables. Group 4 is the same as Group 3 but uses unadjusted energy use. Dashed lines represent *p*-values of 0.5 and 0.1. Whiskers of the boxplots extend to the minimum and maximum *p*-value within 1.5 the interquantile range and outliers are depicted by dots. The black solid lines represent means.

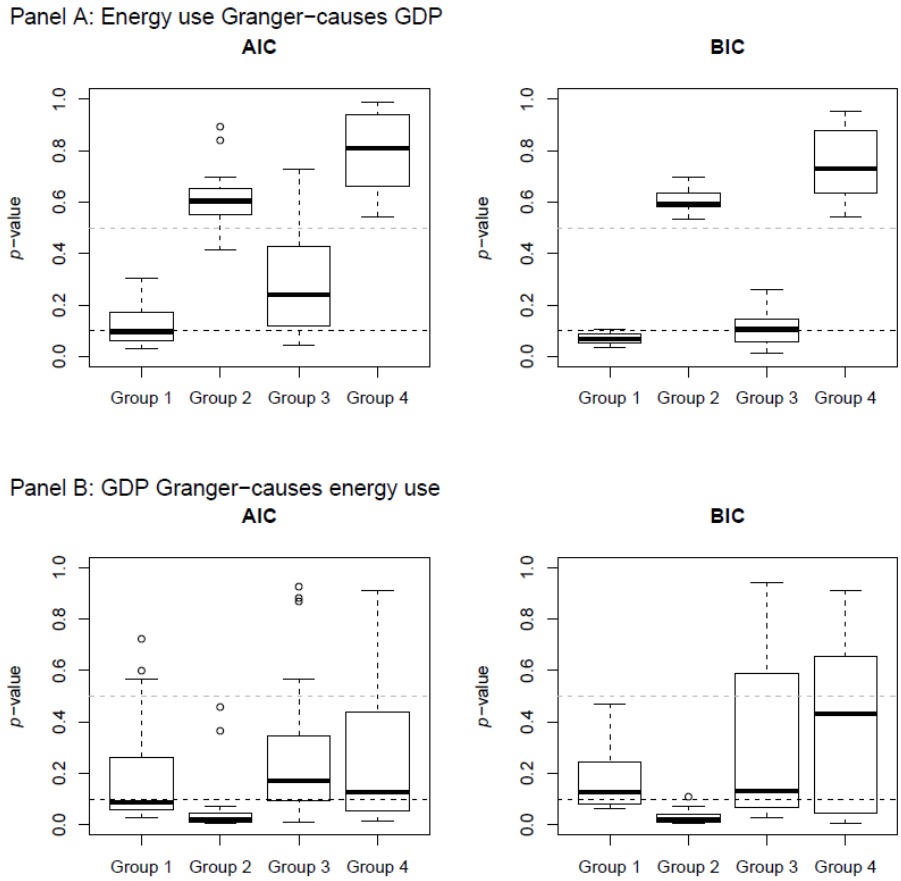

Figure 4.2. Reanalysis of the extended time span 1949-2015 using first differences. Please see caption of Figure 4.1 for further details.

5. Robustness analysis with adjustment for multiple testing

We apply the Benjamini–Hochberg–Yekutieli procedure (Benjamini and Yekutieli, 2011) to adjust *p*-values for multiple testing. We apply the procedure separately for VARs specified by AIC and BIC and separately for VARs estimated with and without structural breaks. Hence, each set of adjusted *p*-values consists of 88 *p*-values (Group 1 to 4).

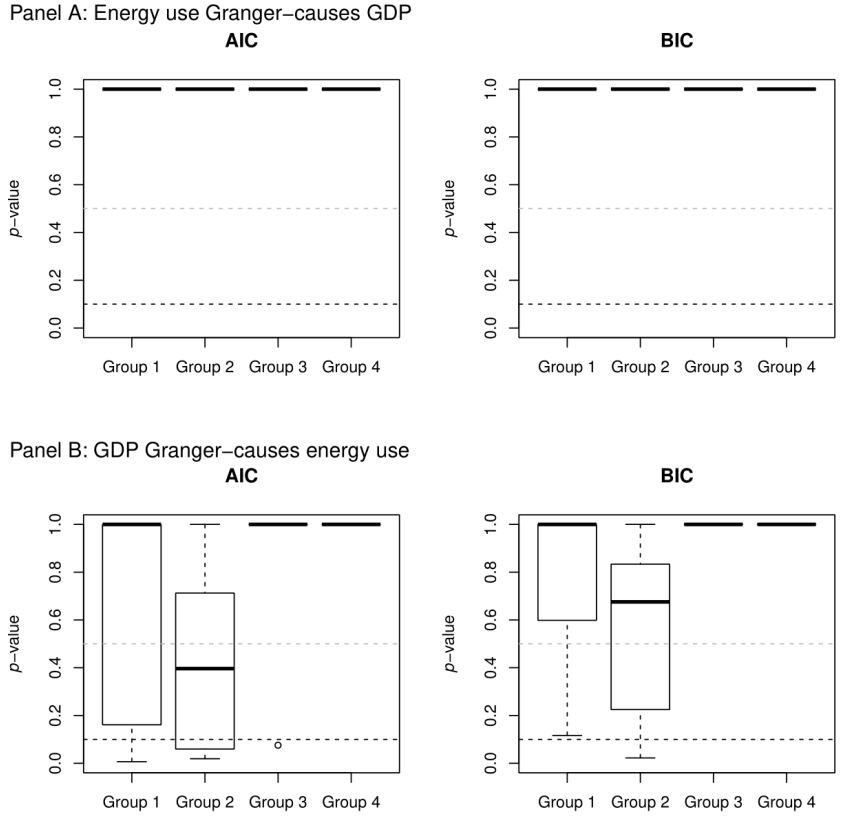

Figure 5.1. Reanalysis of the (almost) original time span 1949-1990. *p*-values adjusted for multiple testing. Please see caption of Figure 4.1 for further details.

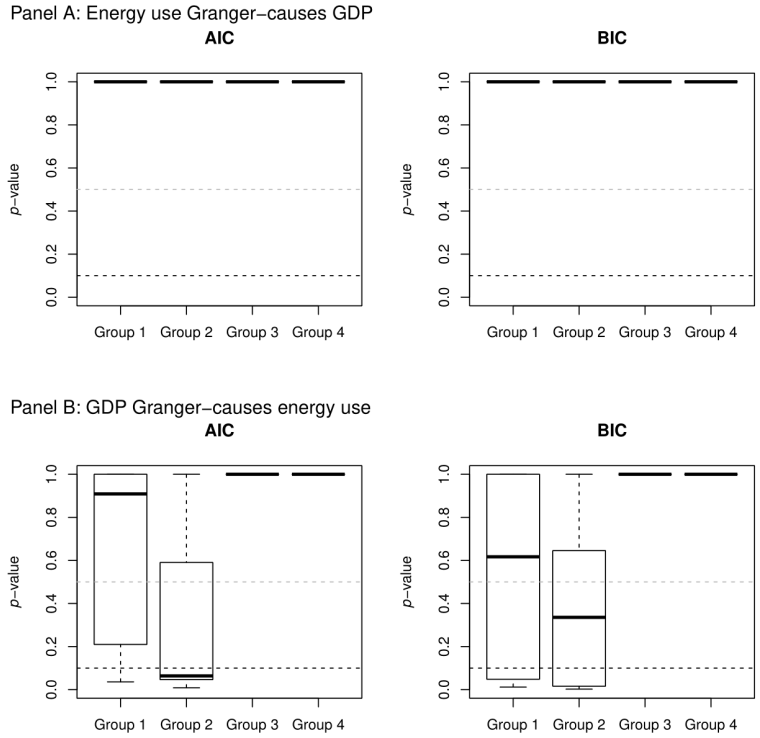

Figure 5.2. Reanalysis of the (almost) original time span 1949-1990 with structural breaks in 1973 and 1979. *p*-values adjusted for multiple testing. Please see caption of Figure 4.1 for further details.

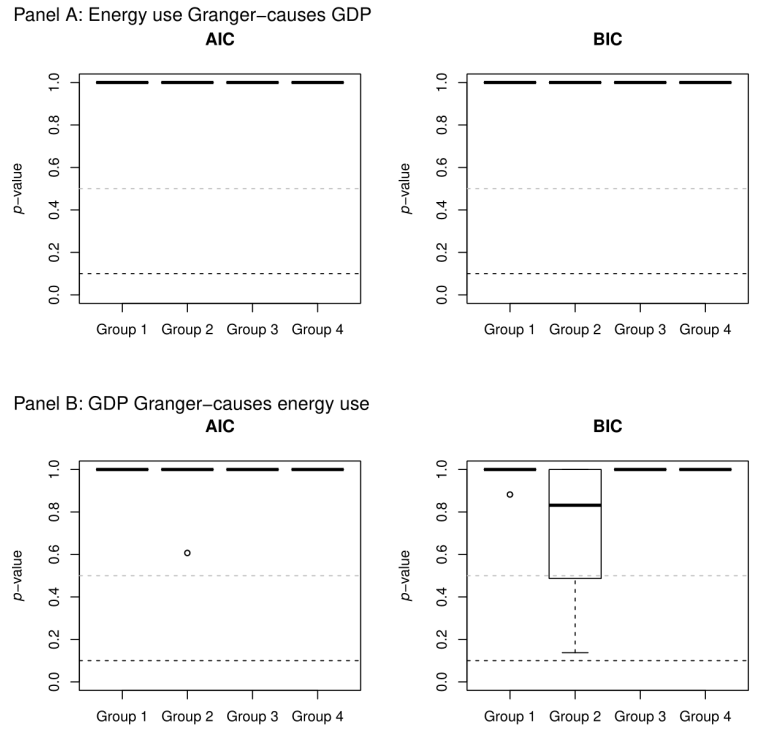

Figure 5.3. Reanalysis of the extended time span 1949-2015. *p*-values adjusted for multiple testing. Please see caption of Figure 4.1 for further details.

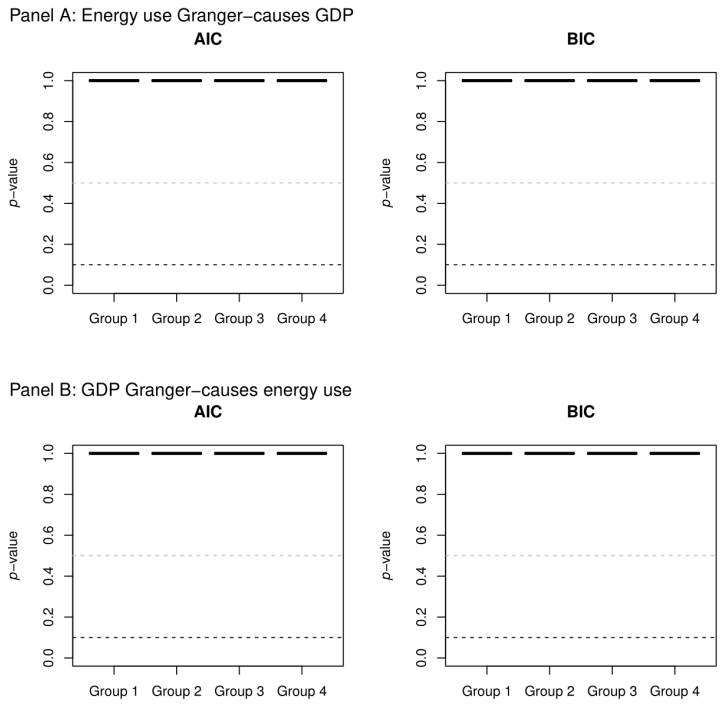

Figure 5.4. Reanalysis of the extended time span 1949-2015 with structural breaks in 1973, 1979 and 2008. *p*-values adjusted for multiple testing. Please see caption of Figure 4.1 for further details.

References

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29(4): 1165–1188.