Lag length selection and *p*-hacking in Granger causality testing: Prevalence and performance of meta-regression models

Stephan B. Bruns*

Department of Economics, University of Göttingen, Humboldtallee 3, 37073 Goettingen, Germany. E-mail: stephan.bruns@uni-goettingen.de.

David I. Stern

Crawford School of Public Policy, The Australian National University, 132 Lennox Crossing, Acton, ACT 2601, Australia. E-mail: david.stern@anu.edu.au.

Abstract

The academic system incentivizes *p*-hacking, where researchers select estimates and statistics with statistically significant *p*-values for publication. We analyze the complete process of Granger causality testing including *p*-hacking using Monte Carlo simulations. If the degrees of freedom of the underlying vector autoregressive model are small to moderate, information criteria tend to overfit the lag length and overfitted vector autoregressive models tend to result in false-positive findings of Granger causality. Researchers may *p*-hack Granger causality tests by estimating multiple vector autoregressive models with different lag lengths and then selecting only those models that reject the null of Granger non-causality for presentation in the final publication. We show that overfitted lag lengths and the corresponding false-positive findings of Granger causality can frequently occur in research designs that are prevalent in empirical macroeconomics. We demonstrate that meta-regression models can control for spuriously significant Granger causality tests due to overfitted lag lengths. Finally, we find evidence that false-positive findings of Granger causality may be prevalent in the large literature that tests for Granger causality between energy use and economic output, while we do not find evidence for a genuine relation between these variables as tested in the literature.

Keywords: Granger causality, *p*-hacking, publication bias, information criteria, meta-analysis, vector autoregression

JEL Codes: C12, C18, C32, Q43

Acknowledgments: We thank Alessio Moneta, participants at the Meta-Analysis in Economic Research Network Colloquium 2013 in Greenwich, participants at the Empirical Workshop on Energy 2014 in Kassel, participants at the IWH-CIREQ macroeconometric workshop 2015 in Halle, and anonymous referees for helpful comments. All remaining errors are ours.

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

1. Introduction

The tendency to selectively publish statistically significant or theory-confirming results may distort the conclusions drawn from published empirical research (Ioannidis, 2005; Glaeser, 2011; Ioannidis and Doucouliagos, 2013). In the case of Granger causality testing, spuriously statistically significant results can be generated if the lag length of the underlying vector autoregressive (VAR) model is overfitted, which is increasingly likely the smaller the sample size. Overfitted lag lengths can occur in standard research designs and they also provide increased opportunities for *p*-hacking. *p*-hacking refers to researchers running many analyses but then only selecting the analyses with statistically significant estimates for the final publication (Simonsohn *et al.*, 2014). In this article, we simulate the complete process of Granger causality testing including *p*-hacking, and we examine how meta-regression models can help identifying genuine Granger causality if the primary literature is distorted by *p*-hacked Granger causality tests.

The current practice of empirical research is largely based on rejecting null hypotheses by finding statistically significant results, usually determined by a *p*-value below 0.05.[1] While misuse and misunderstanding of *p*-values is widespread (Wasserstein and Lazar, 2016), the academic publishing system favors statistically significant results for publication resulting in incentives for individual researchers to search for and select statistically significant results to be presented in submitted articles (Ioannidis, 2005; Glaeser, 2011). Vivalt (2017) provides empirical evidence for this selection in the case of impact evaluation studies by showing that marginally significant estimates are over-represented compared to marginally non-significant estimates. More generally, Brodeur *et al.* (2016) find a lack of *p*-values between 0.10 and 0.25 among more than 50,000 estimates published in the *American Economic Review,* the *Journal of Political Economy,* and the *Quarterly Journal of Economics*. These missing test statistics can be retrieved just below the 0.05 threshold of statistical significance. Reviewing 159 meta-analyses based on more than 60,000 estimates, Ioannidis *et al.* (2016) find that many research fields in empirical economics mainly present statistically significant estimates despite most underlying studies in those fields being underpowered, that is, using sample sizes that are too small to reliably detect the effect of interest. In regression analysis of observational data, *p*-hacking is often based on omitted-variable biases that result from researchers varying the set of control variables included in the regression model. These omitted-variable biases help to

[1] For an overview, see Cumming (2014).

generate statistically significant estimates even in the absence of a genuine effect (Leamer, 1983; Bruns and Ioannidis, 2016; Bruns, 2017).

In Granger causality testing, there is an additional layer of flexibility, as not only the set of control variables but also the lag length of the underlying VAR model needs to be selected. Granger causality test statistics are very sensitive to the lag length chosen for the underlying VAR model (e.g. Zapata and Rambaldi, 1997). Given the importance of this step in Granger causality testing, the choice of lag length is usually based on objective criteria. Frequently used lag length selection criteria are the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). However, these information criteria have a known tendency to overestimate and underestimate, respectively, the true lag length (Nickelsburg, 1985; Lütkepohl, 1985; Ozcicek and McMillin, 1999; Hacker and Hatemi-J, 2008). In the absence of genuine Granger causality, overfitted VAR models also tend to lead to over-rejection of the null hypothesis of Granger non-causality compared to the rejection rate of a VAR model estimated with the true lag length (Zapata and Rambaldi, 1997). *p*-hacking can then be based on selection over various VAR models with different lag lengths. As overfitted lag lengths particularly occur in small samples (Lütkepohl, 2007, pp. 153-157), this source of false-positive findings of Granger causality may be prevalent in macroeconomic research using annual data.

Many approaches have been developed to improve the probability of selecting the correct lag length. These approaches include corrections to the AIC or BIC in small samples (Hurvich and Tsai, 1989). The application of these approaches has, however, been limited and the VARs used in Granger causality testing are usually specified using the standard AIC and BIC, as is mostly the case in the Granger causality literature on energy consumption and economic growth (Bruns *et al*., 2014).

Dealing with false-positive findings of Granger causality due to overfitted lag lengths is important, as researchers are incentivized to *p*-hack, and, as a result, many published Granger causality tests may be spuriously statistically significant. As Cumming (2014) points out, meta-analytical thinking can help deal with biases and improve the reliability and credibility of empirical research. We propose a meta-regression model that synthesizes Granger causality tests from many primary studies to help identify the presence or absence of genuine Granger causality while controlling for potential biases.

Meta-regression analysis in economics was originally proposed to explain the variation in empirical findings (Stanley and Jarrell, 1989). Meta-regression analysis was further developed to identify genuine empirical effects while controlling for p-hacking based on sampling errors (Stanley, 2008; Bruns, 2017). These approaches use the concept of statistical power to determine if a genuine effect exists across a sample of primary studies. If there is a genuine effect, test statistics from the primary studies, such as the t-statistic for a regression coefficient, should increase with the degrees of freedom used in the underlying primary estimates, whereas in the absence of a genuine effect the test statistics should be unrelated to the degrees of freedom.

Meta-regression models have been primarily developed for the synthesis of single regression coefficients, which consist of a point estimate and a standard error. The standard approach to testing for a genuine effect is to regress the ratio of the estimated coefficient and its standard error on a constant, the inverse of the standard error, and control variables. But Granger causality tests are usually F or χ^2-distributed test statistics derived from restricting multiple coefficients in a model. So, both this and the potential false-positive findings of Granger causality due to overfitted lag lengths need to be taken into account in using meta-regression models to analyze Granger causality test statistics.

Our meta-regression model for Granger causality tests regresses the probit-transformed p-values of the original Granger causality test statistics on a constant, the square root of the degrees of freedom in the primary regressions, and the selected lag length from the primary studies. Using Monte Carlo simulations, we show that overfitted lag lengths and the corresponding prevalence of false-positive findings of Granger non-causality occur in many scenarios that are likely to be prevalent in macroeconomics. We also simulate empirical literatures that are distorted by p-hacking based on overfitting the lag length or exploiting sampling errors. Our results reveal that p-hacking based on overfitting lag lengths may result in empirical literatures that are characterized by false-positive findings of Granger causality. Our simulation results also show that our proposed meta-regression model can help identify whether statistically significant Granger causality tests in published studies stem from genuine Granger causality or from p-hacked Granger causality tests.

We use the large literature that tests for Granger causality between energy use and economic output to evaluate how common spuriously significant Granger causality tests due to overfitted lag lengths are. We show that the excess significance in this literature can be explained by overfitted lag lengths rather than the presence of linear Granger causality between energy use and economic output. These findings highlight how as a result of overfitted lag lengths a literature can appear to provide evidence for Granger causality when actually Granger causality appears to be absent.

Section 2 of the paper discusses testing for Granger causality, overfitted lag lengths, p-hacking, and the meta-regression models. Section 3 describes the designs of the Monte-Carlo simulations and presents the results. Section 4 investigates the literature on energy use and economic output. Section 5 discusses the findings and Section 6 concludes.

2. Meta-regression analysis of Granger causality tests

2.1. Testing for Granger causality

Granger (1969) introduced a concept of causality that is based on the idea that the future cannot cause the past. Assuming stationarity, a variable X is said to Granger-cause a variable Y if past values of X help explain the current value of Y given past values of Y and all other relevant past information U. Let U' be the set of all information up to and including period t-1 apart from observations on X. If $E(Y|U) \neq E(Y|U')$, then X causes Y (Granger, 1988). In applied econometrics, the whole universe of information is not available, and the functional form is usually assumed to be linear. Hence, in practice, Granger causality tests are usually based on improved linear prediction within a specific model (Lütkepohl, 2007, pp. 41-43).

As we focus our analysis on overfitting the lag length and p-hacking in Granger causality testing, we concentrate on the Granger causality testing procedure of Toda and Yamamoto (1995) that avoids the potential occurrence of additional biases due to pre-testing the order of integration or cointegration. This testing procedure is frequently applied in the energy-growth literature. Toda and Yamamoto (1995) show that if a VAR in levels is augmented by the number of lags equal to the highest degree of integration, a Wald test that does not restrict the augmenting lags is asymptotically χ^2-distributed irrespective of the order of integration and cointegration. Hence, we can test for Granger causality by estimating the following VAR (ignoring any deterministic components) and testing restrictions on its coefficients:

$$Y_t = \Pi_1 Y_{t-1} + \cdots + \Pi_p Y_{t-p} + \Pi_{p+1} Y_{t-p-1} + \cdots + \Pi_{p+d_{max}} Y_{t-p-d_{max}} + \varepsilon_t \quad (1)$$

where Y_t is a $k \times 1$ vector of variables, Π_i is a $k \times k$ matrix of coefficients, ε_t is a $k \times 1$ vector of errors, p denotes the lag length and d_{max} is the maximal order of integration. We can test for Granger causality from $Y^{(a)}$ to $Y^{(b)}$, where the superscripts denote two individual variables in Y_t, using H_0: $\Pi_1^{ab} = \Pi_2^{ab} \dots = \Pi_p^{ab} = 0$, where the superscripts denote the ath column and bth row of Π_i. Stacking the coefficient matrices as $\Pi = vec[\Pi_1, \Pi_2, \dots, \Pi_{p+d_{max}}]$ and letting R be the matrix of restrictions so that $R\Pi = vec[\Pi_1^{ab}, \Pi_2^{ab}, \dots, \Pi_p^{ab}]$, then H_0: $R\Pi = 0$ can be tested by a Wald test:

$$W_p = (R\widehat{\Pi})'\left[R\widehat{\Sigma}_p R'\right]^{-1} R\widehat{\Pi} \qquad (2)$$

where W is asymptotically χ_p^2 distributed with p degrees of freedom, $\widehat{\Sigma}_p$ is the estimated covariance matrix of (1) and $\widehat{\Pi}$ is the estimate of Π.

2.2. Overfitted lag lengths and *p*-hacking

It is common to estimate many different VAR models in the research process. *p*-hacking refers to the selective presentation of those VAR models that guarantee a *p*-value below the typical thresholds of statistical significance for the Granger causality test of interest, while a potentially large number of estimated VAR models remain unreported (Simonsohn *et al.*, 2014). For example, *p*-hacking can be based on omitted-variable biases if the researcher varies the set of control variables until a *p*-value below the desired significance level is obtained (see for example, Leamer, 1983; Bruns and Ioannidis, 2016; Bruns, 2017). But *p*-hacking can be also based on sampling errors if researchers vary the sample by, for example, changing the years and/or countries included in a panel data set (Bruns, 2017).[2]

In Granger causality testing, an additional layer of flexibility in the research design is introduced by the need to specify a lag length for the underlying VAR model. The choice of the lag length in VAR models is mainly an empirical question, as economic theory is usually not very specific about the temporal dimension of economic dynamics. Although there are various methods for determining the lag length, information criteria, such as the AIC and BIC, are most commonly used. It is well known that the BIC is consistent in estimating the correct lag length while the AIC overfits (Lütkepohl, 2007, pp. 146). However, these asymptotic properties may have little relevance for lag length selection in economic time series. In contrast to the high frequency data widespread in finance, macroeconomic time series usually consist

[2] Variation of the set of analyzed countries or years may of course also change the effect that is estimated if there is heterogeneity in the effect of interest. Thus, *p*-hacking based on sampling errors may easily become *p*-hacking based on selection from heterogeneity in the effect of interest.

of a few decades of quarterly or annual data. Hence, there is usually a small to moderate number of observations.

Accordingly, it is the performance of information criteria in small and moderate sample sizes that matters in applied macroeconometrics. Although the exact frequency with which the correct lag length (p^*) is chosen may vary with respect to the specific DGP, systematic patterns can be identified when information criteria are used (for an overview see Lütkepohl, 2007, pp. 146-157). The probability to overfit a VAR(p^*) model by h lags is given by

$$P[IC(p^*) > IC(p^* + h)] = P\left[ln\left|\hat{\Sigma}_{p^*}\right| - ln\left|\hat{\Sigma}_{p^*+h}\right| > \frac{c_T\, p^*\, q^2\, h}{T}\right] \quad (3)$$

where IC is the information criterion, $\hat{\Sigma}_{p^*}$ is the estimated covariance matrix of the VAR(p^*) model, T is the number of observations, q is the dimension of the VAR model, and c_T is a penalty term. If there are few degrees of freedom, the sampling variability of $\hat{\Sigma}$ will be large. As a result, the variance of $ln\left|\hat{\Sigma}_{p^*}\right| - ln\left|\hat{\Sigma}_{p^*+h}\right|$ can become large while the penalty term is not affected by sampling variability. Accordingly, the probability of overfitting is higher, the lower the number of degrees of freedom. Moreover, given that the AIC uses $c_T = 2$ and the BIC uses $c_T = \ln(T)$, the penalty term is systematically larger for the BIC than for AIC if $T > 7$. Therefore, the probability that the IC suggests an overfitted VAR is larger for the AIC than for the BIC. Analogously, the probability of underfitting a $VAR(p^*)$ model by h lags is given by

$$P[IC(p^*) > IC(p^* - h)] = P\left[ln\left|\hat{\Sigma}_{p^*-h}\right| - ln\left|\hat{\Sigma}_{p^*}\right| < \frac{c_T\, p^*\, q^2\, h}{T}\right]. \quad (4)$$

The potentially large variance of $ln\left|\hat{\Sigma}_{p^*-h}\right| - ln\left|\hat{\Sigma}_{p^*}\right|$ due to sampling variability for low degrees of freedom implies that there is also an increased probability of underfitting. As the penalty term is larger for the BIC, the probability of underfitting is larger for the BIC than for the AIC. These patterns have been shown in simulations for a variety of DGPs including VARs with high lag lengths (Nickelsburg, 1985) and low lag lengths (Lütkepohl, 1985) as well as stable and unstable VARs under situations with homoscedasticity or ARCH (Hacker and Hatemi-J, 2008) and symmetric or asymmetric lag lengths (Ozcicek and McMillin, 2010).

VAR models with overfitted lag lengths tend to over-reject the null hypotheses of Granger non-causality compared to the rejection rate of a VAR model estimated with the true lag length (Zapata and Rambaldi, 1997). As a result, overfitted VAR models lead to an increased rate of false-positive findings of Granger causality. p-hacking based on overfitted lag lengths occurs if researchers use overfitted lag lengths to produce statistically significant estimates, which they then select for presentation in the final paper.

It is important to emphasize that published articles that use overfitted VAR models with spuriously significant Granger causality tests are not necessarily the result of p-hacking. The use of information criteria is a standard approach to specify the lag length suggested in many econometric textbooks, and overfitted lag lengths can also occur even if researchers do not select a few VAR models from a large set of estimated VAR models for the final publication.

2.3. Meta-regression model for Granger causality tests

The following basic meta-regression model for Granger causality test statistics (Bruns *et al.*, 2014) aims to identify whether there is genuine Granger causality in the presence of p-hacking based on sampling errors but not based on overfitted lag lengths:

$$z_i^{gc} = \alpha_B^{gc} + \beta_B^{gc} \sqrt{df_i} + \varepsilon_i^{gc} \tag{5}$$

where df_i is the degrees of freedom of a single equation of the VAR used in primary study i and $z_i^{gc} = \Phi^{-1}(1 - \pi_i^{gc})$, where π_i^{gc} is the p-value of study i and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution, also known as probit. Larger values of z_i^{gc} indicate smaller p-values and, consequently, higher levels of statistical significance. The direction of Granger causality tested is given by $g = 1, \dots, q$ denoting the equation in the VAR and $c = 1, \dots, q$ denoting the variable in equation g so that, for example, $g = 1$ and $c = 2$ represents Granger causality from the second variable to the dependent variable in the first equation of the VAR.[3]

If there is no genuine effect, the probit transformation of the p-values results in a normally distributed dependent variable with mean zero. Hence ε_i^{gc} has desirable properties for a regression residual. In the presence of genuine Granger causality, Toda and Yamamoto's Granger Causality test statistic follows a non-central χ^2-distribution and the level of statistical significance increases as df_i increases ($\beta_B^{gc} > 0$). Conversely, in the absence of genuine Granger causality, df_i should be unrelated to the levels of statistical significance.[4]

[3] This basic model may be augmented by other control variables and interactions between the controls and the degrees of freedom variable in actual applications – see Section 4 of this article or Bruns *et al.* (2014) for more details.

[4] Please note that this only holds if the VAR model is correctly specified and, for example, omitted-variable biases are absent. We discuss this in the empirical application in Section 4.

In sampling error-based p-hacking, large estimates of the VAR coefficients are required to achieve statistical significance when there are few degrees of freedom, whereas smaller estimates of the VAR coefficients are sufficient when there are many degrees of freedom. Hence, the p-values will be unrelated to the degrees of freedom if the primary literature exclusively consists of statistically significant results generated by using sampling errors. Simulations show that meta-regression models of this type can control for this type of p-hacking (Stanley, 2008; Bruns, 2017) and, thus, $H_0: \beta_B^{gc} \leq 0$ tests for the presence of genuine Granger causality. Figure 1 shows how the meta-regression model would behave in three different idealized situations.

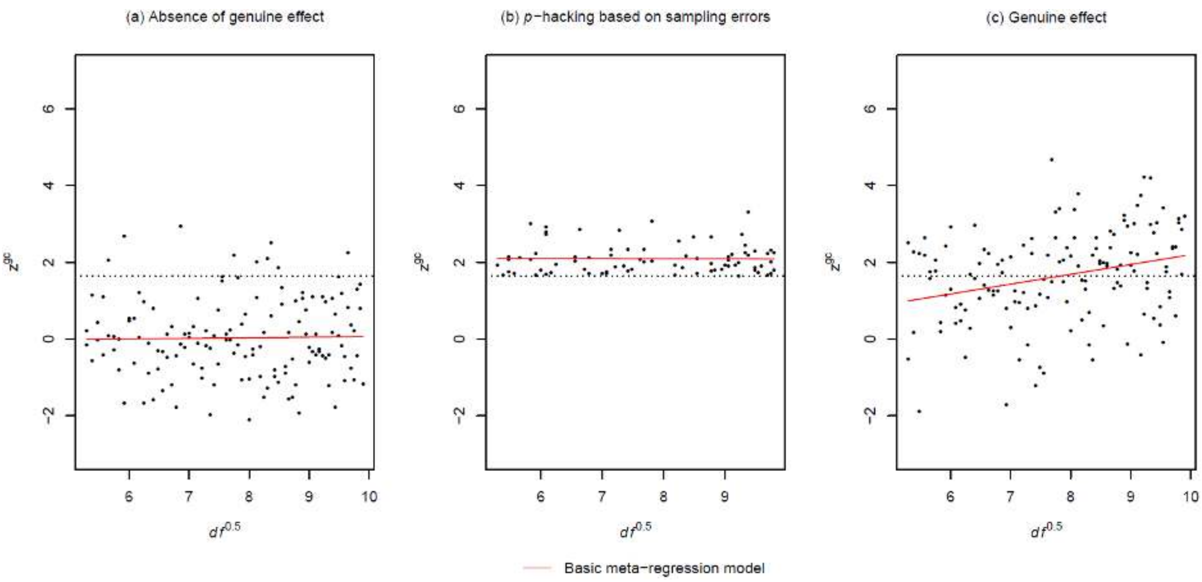

Fig. 1 Properties of the basic meta-regression model are shown. Each graph is a hypothetical illustration of the relationship between probit-transformed p-values and $\sqrt{df}$ in the following three different situations: (a) in the absence of genuine Granger causality, (b) in the absence of genuine Granger causality but with p-hacking based on sampling errors, and (c) in the presence of genuine Granger causality. The dotted line indicates the 0.05 significance level ($z^{gc} = 1.64$). Data points above this line are statistically significant and data points below this line are statistically non-significant. The red solid line illustrates the fit of the basic meta-regression model.

As discussed in Section 2.2, overfitting the lag length might be used to consciously or unconsciously find statistically significant Granger causality tests. Meta-regression analysis can help to identify the presence of genuine Granger causality if spuriously significant Granger causality tests due to overfitted lag lengths are present in the literature. Overfitted lag lengths and the corresponding over-rejection of Granger non-causality leads to large values of z^{gc} compared to the values of z^{gc} that we can expect for models estimated with the true lag length and these large values of z^{gc} are more common in small samples. Therefore, we can expect that β_B^{gc} is biased downwards compared to the true relation between z^{gc} and $\sqrt{df}$. This downward bias in β_B^{gc} reduces the power of the basic meta-regression model. We suggest controlling for the underlying lag length of the VAR model in the meta-regression model to account for this source of bias:

$$z_i^{gc} = \alpha_E^{gc} + \beta_E^{gc} \sqrt{df_i} + \gamma^{gc} p_i + v_i^{gc} . \tag{6}$$

We refer to this model as the extended meta-regression model for Granger causality tests.[5] In the presence of p-hacking based on exploiting sampling errors and overfitting the lag lengths, there is still evidence for genuine Granger causality if we can reject $H_0: \beta_E^{gc} \leq 0$. Fig. 2 illustrates for idealized data the expected behavior of the two meta-regression models in the presence of overfitted lag lengths and the corresponding over-rejection of Granger non-causality. In the absence of a genuine effect, the regression slope for the basic meta-regression model is usually negative, while the coefficient of $\sqrt{df}$ is zero for the extended meta-regression model. Hence, $\beta_B^{gc} < 0$ may be used as an indication that overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality are present in the literature. In the presence of a genuine effect, both β_B^{gc} and β_E^{gc} are positive but β_E^{gc} is larger indicating that overfitted lag lengths reduce the power of the basic meta-regression model compared to the extended meta-regression model.[6]

[5] While Toda and Yamamoto test statistics tend to underreject if the VAR model is underfitted and to overreject if the VAR model is overfitted, this is not generally the case for all Granger causality test procedures (Zapata and Rambaldi, 1997). Therefore, one can consider using dummy variables for each lag length in an extended meta-regression model rather than a continuous variable if other types of Granger causality tests are analyzed.

[6] Note that if genuine Granger causality is present, overrejection of the null of Granger non-causality compared to a model with the true lag length is not common to all types of Granger causality tests (Zapata and Rambaldi, 1997).

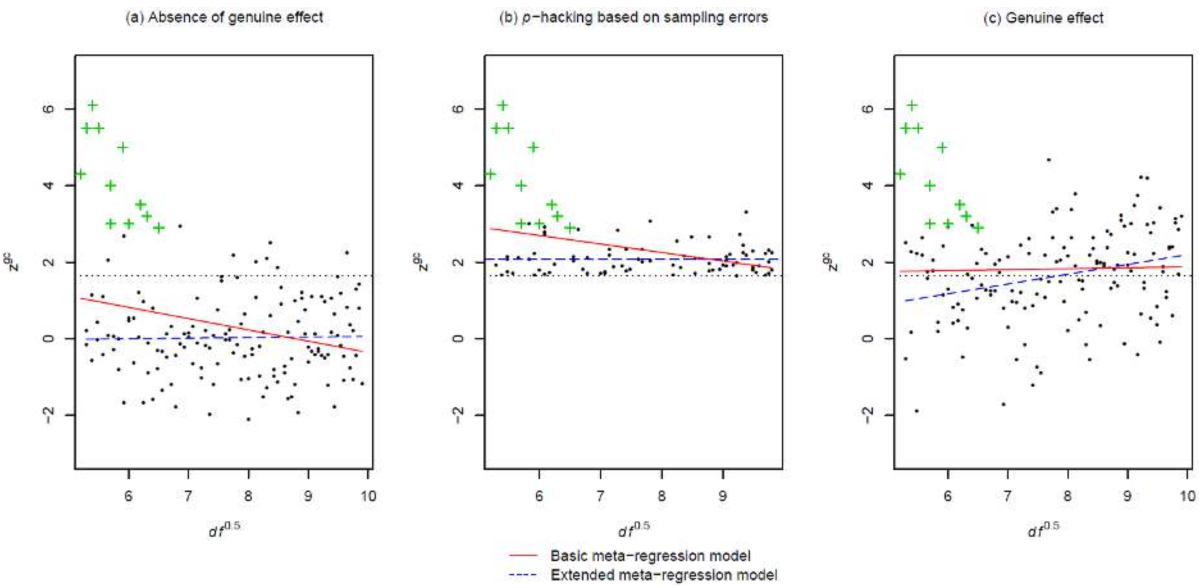

Fig. 2 Properties of both the basic and extended meta-regression model are shown in the presence of overfitted lag lengths. The green crosses represent Granger causality test statistics that are statistically significant due to overfitted lag lengths. The red solid line represents the basic meta-regression model and the blue dashed line represents the extended meta-regression model that controls for the lag lengths. Please see the caption of Fig. 1 for additional information.

3. Monte Carlo simulation

3.1. Design

3.1.1. No *p*-hacking

First, we analyze how prevalent overfitted lag lengths and the corresponding false-positive findings of Granger causality would be if the authors of primary studies use standard research designs and do not engage in any *p*-hacking. We then examine how well the basic and extended meta-regression models perform in this case.

For each simulated meta-regression analysis, we generate $i = 1, \ldots, s$ underlying primary studies with meta-analysis sample sizes s = 10, 20, 40, 80. The sample size of each primary study, n_i, is selected by first drawing a number from a gamma distribution with scale parameter $\frac{\sigma^2}{(\mu-30)}$ and shape parameter $\frac{(\mu-30)^2}{\sigma^2}$ to which we then add 30 and round to the nearest integer. This allows us to vary the mean μ and the variance σ^2 independently and it ensures that $n_i = 30$ is the smallest primary sample size. We consider $\mu = 35, 40, 50, 60$ and $\sigma^2 = 25, 100, 225$ to mirror a wide span of primary sample size distributions ranging from rather small primary sample sizes typical for annual data in macroeconomics to larger primary sample sizes that are more likely to be present in quarterly or monthly data in macroeconomics.

Annual macroeconomic time series often start in 1970 but may start earlier or later. For example, most series in the *World Bank Development Indicators* start in 1980. If the meta-analyst considers primary studies using annual data published in the last 15 years, the primary sample sizes may range between 30 and 55, though some primary studies may use time series for specific countries that are substantially longer. Such a distribution is mirrored by $\mu = 40$ and $\sigma^2 = 100$ illustrated in Fig. 3. The 10% (90%) quantile is 31 (53) and the distribution is right skewed and allows for the presence of some large primary sample sizes. A similar but more symmetric distribution with less probability mass on larger primary sample sizes is given by $\mu = 40$ and $\sigma^2 = 25$. This distribution is also illustrated in Fig. 3 and provides a 10% (90%) quantile of 34 (47). Quarterly time series provide more frequent observations but are often available for fewer years. If the meta-analyst considers studies published in the last 15 years, the primary sample sizes may range between 40 and 80 with some larger samples. Fig. 3 illustrates how these primary sample sizes are mirrored by $\mu = 60$ and $\sigma^2 = 15^2$ leading to a distribution with a 10% (90%) quantile of 43 (80).

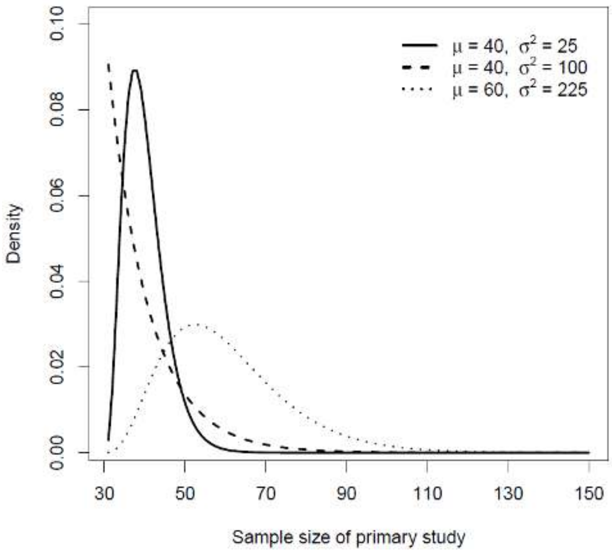

Fig. 3 Distributions of primary sample sizes for different combinations of μ and σ^2 are shown.

We generate data for the primary studies using four DGPs (Table 1). All four DGPs have a true lag length of three (p = 3) so that we can illustrate both underfitting and overfitting. Following

Zapata and Rambaldi (1997), all DGPs imply that X causes Y but not *vice versa*, which allows us to evaluate the size and power of the meta-regression models using the same DGP.

DGP1 is a bivariate VAR process with two unit roots. We set the two coefficients on the diagonal of each matrix equal in order to focus on the ability of the meta-regression models to detect the causal effect, which is determined by the off-diagonal coefficients. DGP1a and DGP1b only differ with respect to the strength of the causal effect with DGP1b having a larger casual effect, allowing us to evaluate the performance of meta-regression models for different sizes of causal effects. DGP2 is a bivariate VAR process with one unit root so that the model is cointegrated. DGP2a and DGP2b only differ with respect to the causal effect with DGP2b having a larger causal effect. The residuals are modeled as $\epsilon_t \sim N(0, \Omega)$ where $\Omega = \mathrm{I}$ or $\Omega = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ so that there are two cases for each DGP.

For each primary study i, we draw three starting values for X and Y from standard normal distributions and generate $n_i + 50$ observations using one of the DGPs. Afterwards we delete the first 50 observations to reduce dependence on the starting values.

Each primary study determines the optimal lag length ($p = 1, \dots, 5$) for the VAR in levels using either the AIC or BIC. Subsequently, the lag length is augmented with the maximum order of integration of one. As a result, the minimum number of degrees of freedom that a primary study can have is 11. Finally, each primary study applies a Wald test to the lags of the independent variable ignoring the additional augmenting lag which produces Granger causality tests for X causes Y and Y causes X for each DGP.

We apply the basic and extended meta-regression model to the s primary studies and evaluate their size and power in identifying genuine Granger causality. We use 1000 iterations for each of the 768 scenarios ($\#s * \#\mu * \#\sigma^2 * \#DGP * \#IC * \#\Omega$). We use the same simulated data set to also evaluate how prevalent overfitted lag lengths and the corresponding over-rejections of the null of Granger non-causality are for a given primary sample size distribution (μ and σ^2). There are 150,000 Granger causality test statistics for each combination of DGP, IC, and Ω.

Table 1 Overview of data-generating processes

Name	Vector autoregressive model
DGP1a	$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 1.5 & 0.4 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.2 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.2 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$
DGP1b	$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 1.5 & 0.8 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$
DGP2a	$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 1.5 & 0.4 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.5 & 0.2 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.2 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$
DGP2b	$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 1.5 & 0.8 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$

3.1.2. Theory-confirmation bias

We also examine the case where researchers search for theory-confirming and statistically significant Granger causality tests. Suppose theoretical considerations suggest that there is a causal effect from Y to X, when in fact causality is actually absent in this direction and may or may not be present from X to Y. If these theoretical considerations dominate the empirical literature, researchers may *p*-hack to confirm these theoretical presumptions by consciously or unconsciously overfitting the lag length.

We again generate primary sample sizes, n_i, as described in section 3.1.1. Each study tests for Granger causality from Y to X based on a VAR model that is specified using the AIC and a VAR model that is specified using the BIC. Each primary study then selects for publication the test of Granger causality from Y to X that is more statistically significant. Moreover, we consider that $h\%$ (where $h = 0, 25, 50, 75,$ or 100) of the primary studies not only select the more statistically significant result for causality from Y to X from the AIC- and BIC-specified models, but they also search further samples of data (e.g. other countries or time periods) until they find Granger causality from Y to X that is statistically significant at the 0.05 level. We simulate this by generating further samples from the relevant DGP and fitting VAR models to them using the AIC and BIC until the more statistically significant Granger causality test from

Y to X is statistically significant at the 0.05 level. This gives further opportunities to generate apparently significant results due to sampling errors and overfitted lag lengths.

As a result, the primary literature is composed of h% primary studies with statistically significant Granger causality tests from Y to X due to p-hacking based on exploiting sampling errors and overfitting lag lengths. The remaining $(1-h)$% primary studies only search for the desired result by specifying the lag length of the VAR model using the AIC and BIC and selecting the more significant result in the direction of Y to X. If these $(1-h)$% primary studies do not obtain a statistically significant and theory-confirming result, they publish their findings anyway. The outcome is an empirical literature that provides systematic support for a false theory that increases with h. We use 1000 iterations for each of the 1920 scenarios ($\#s * \#\mu * \#\sigma^2 * \#DGP * \#h * \#\Omega$). As we did in the case without p-hacking, we use this simulated data to also evaluate the prevalence of overfitted lag lengths and the corresponding over-rejections of the null of Granger non-causality for a given primary sample size distribution (μ and σ^2).

3.2. Results

3.2.1. No p-hacking

Our results show that overfitting occurs frequently for the AIC, whereas the BIC tends to underfit the true lag length. Both the AIC and the BIC overfit when the degrees of freedom are small. In the presence of genuine Granger causality (i.e. tests of X causes Y), the p-values of the Granger causality tests are largely below the nominal significance level of 0.05. In the absence of genuine Granger causality (i.e. tests of Y causes X), the p-values of the Granger causality tests tend to become smaller – i.e. more statistically significant – as the lag length increases. Overfitted VAR models have p-value distributions with a smaller mean than the VAR model with the true lag length of three mirroring over-rejection of the null of Granger non-causality. Underfitted VAR models have p-value distributions with a larger mean compared to the VAR model with the true lag length mirroring under-rejection. Fig. 4 illustrates these findings for DGP2a with $\Omega = \mathrm{I}$, and Appendix A1 shows the results for the remaining DGPs. The simulation reveals that, especially when the AIC is used, overfitted lag lengths and over-rejection of the null of Granger non-causality occur frequently in a variety of scenarios that mirror actual research in empirical macroeconomics.

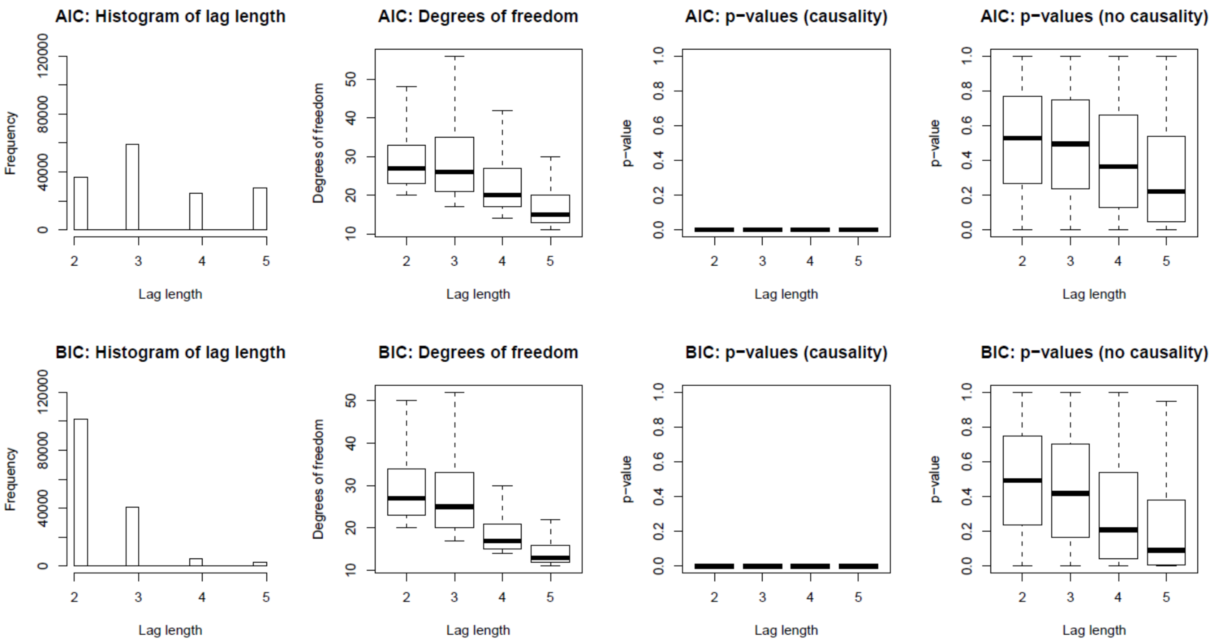

Fig. 4 Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP2a. The first column shows the histograms of selected lag lengths in simulated primary studies across all meta-sample sizes ($s = 10, 20, 40, 80$) resulting in 150,000 observations using a primary sample size distribution with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = \mathrm{I}$. The second column presents the boxplots of degrees of freedom by lag length. The box represents the interquartile range and the whiskers extend to the largest data point within 1.5 times the interquartile range. The third column shows the boxplots of p-values in simulated primary studies in the presence of Granger causality, whereas the fourth column presents the boxplots of p-values in the absence of Granger causality. A lag length of one was selected for less than 0.1% of primary studies and these findings are not reported.

Fig. 5 shows how the type I errors of both meta-regression models vary with the meta-sample size for DGP2a and DGP2b (the cointegrated DGP). The type I errors of the basic meta-regression model are mostly smaller than the type I errors of the extended meta-regression model due to the downward bias of β_B^{gc}. The type I errors of the extended meta-regression model are largely below but close to the nominal significance level of 0.05. This shows that β_E^{gc} is still biased downwards. DGP1 shows the same patterns as DGP2 (See Appendix A1 for DGP1).

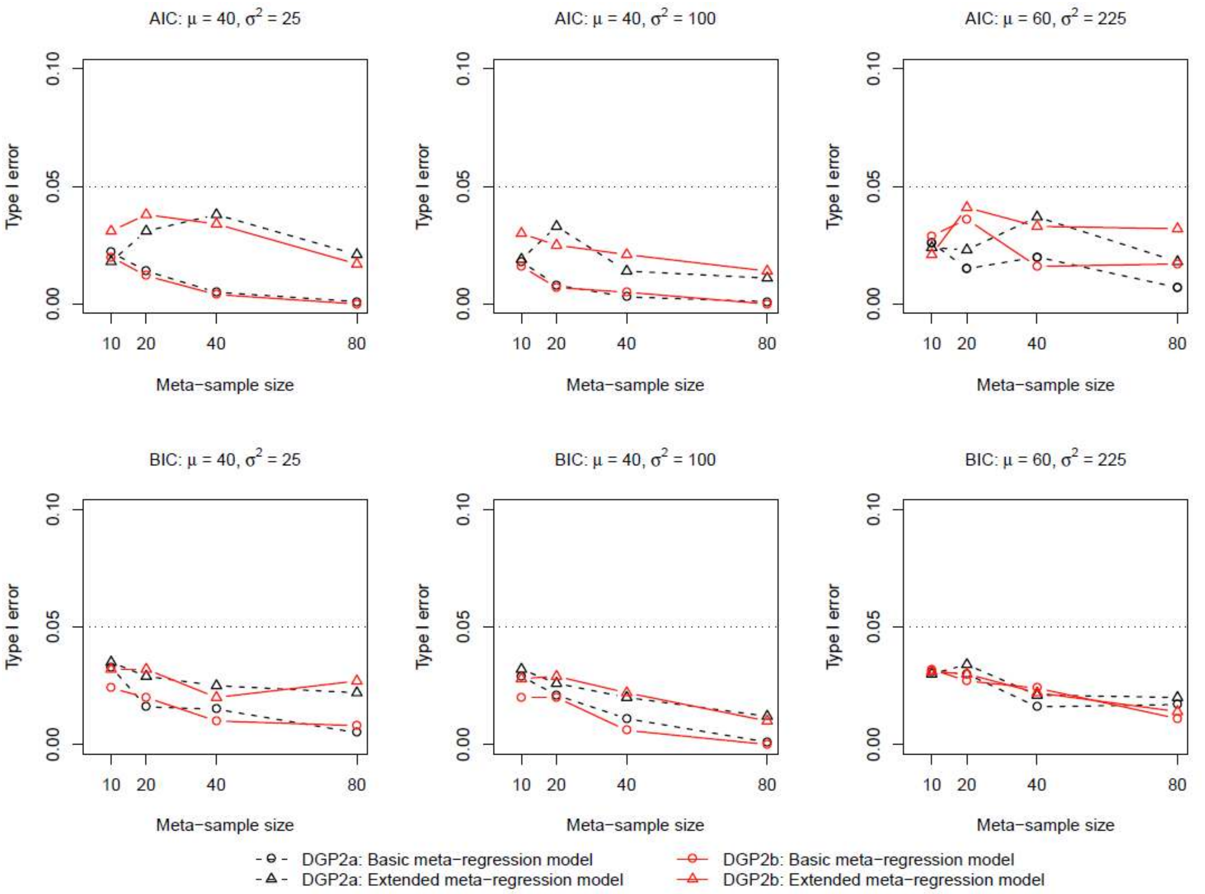

Fig. 5 Type I errors of both the basic and extended meta-regression models for DGP2a and DGP2b are shown. Type I errors of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = I$ are reported if the AIC (upper row) or the BIC (lower row) is used for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Fig. 6 shows the power of both meta-regression models in identifying genuine Granger causality in relation to the meta-sample size for DGP2a and DGP2b. For very small meta-sample sizes, the basic model can have higher power than the extended model as adding the lag length as a control variable reduces the degrees of freedom of the meta-regression model. However, as the meta-analysis sample size increases, the power of the extended model increases more strongly than the power of the basic model. The difference between the basic and extended meta-regression model is especially large for low primary study sample size means, as the probability of overfitting is larger in small samples. The difference between these two meta-regression models diminishes as the variance, σ^2, of the primary sample sizes or the mean, μ, become larger. The difference is higher if the actual causal effect is small, as the downward bias of β_B^{gc} in the basic model results more easily in acceptance of $H_0\colon \beta_B^{gc} \leq 0$ even though genuine Granger causality is present. Using the BIC results in a larger difference between the basic and extended meta-regression models than using the AIC, though overfitted lag lengths are actually more prevalent for the AIC. The reason is that the use of BIC leads to overfitted VAR models with exceptionally small degrees of freedom. The difference between the two meta-regression models decreases if the VAR errors are correlated.

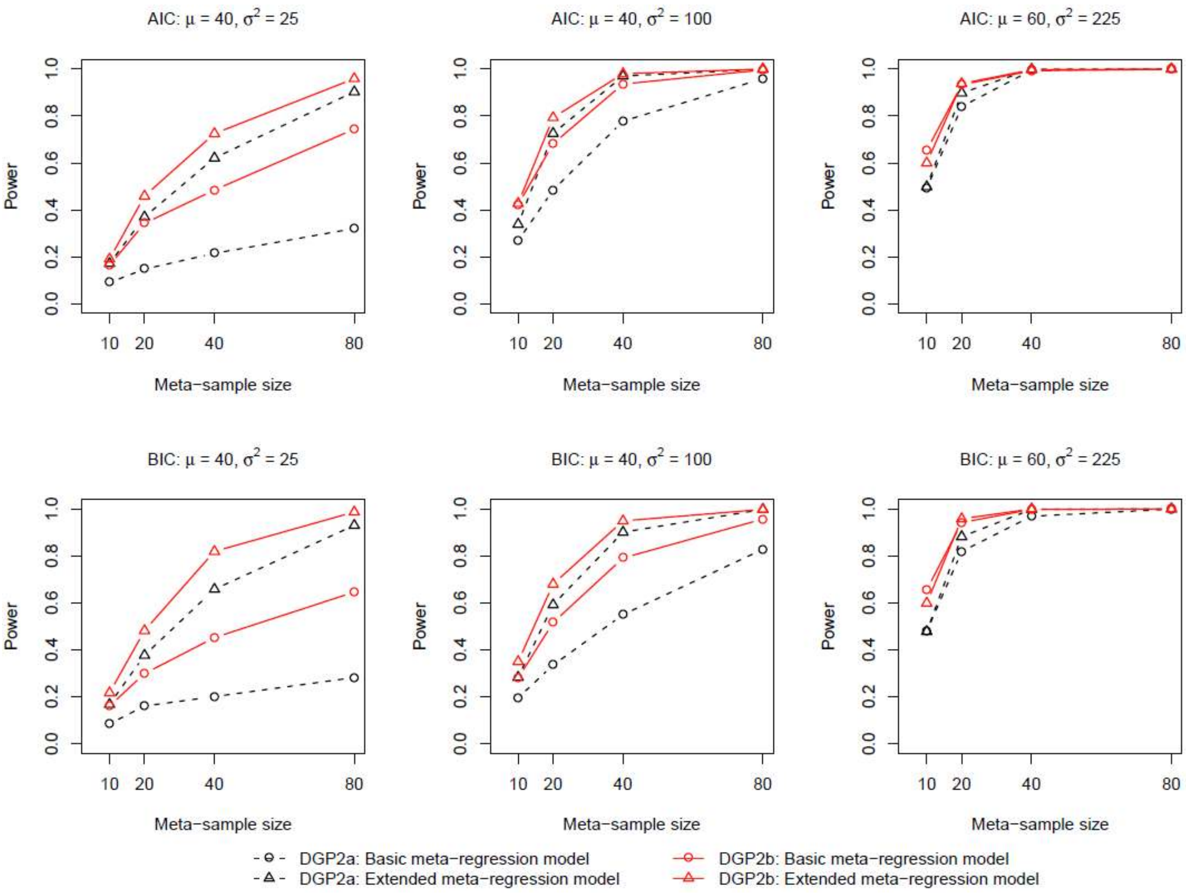

Fig. 6 Power of both the basic and extended meta-regression models for DGP2a and DGP2b are shown. Power curves of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = \mathrm{I}$ are reported if the AIC (upper row) or the BIC (lower row) is used for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Power increases if the primary sample size distribution becomes larger or if the actual causal effect is larger, and it decreases if the VAR errors are correlated across equations. DGP1 shows the same patterns as DGP2 but with systematically smaller power revealing cointegration as an important determinant of power (See Appendix A1 for DGP1).[7]

[7] As an anonymous reviewer pointed out, economic time series may often be highly persistent but stationary (Nelson and Plosser, 1982). We also considered a VAR process that is stationary, but its two largest characteristic roots are 0.95:

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 0.95 & -0.475 \\ 0 & 0.95 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} 0.25 & -0.125 \\ 0 & 0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.2375 & 0.11875 \\ 0 & -0.2375 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

3.2.2. Theory-confirmation bias

In the second simulation design, the primary study authors search for statistically significant and theory-confirming results, that is Granger causality from Y to X, where genuine Granger causality is actually absent. Fig. 7 shows that overfitted VAR models are more prevalent in this case, indicating that p-hacking is based on both overfitted lag lengths and sampling errors. A large amount of excess significance is present for Y causes X, indicating how distorted an empirical literature could become.[8]

The type I errors of both meta-regression models are again well below the nominal significance level of 0.05. Fig. 8 shows how they vary with the degree of p-hacking for DGP2a and DGP2b. Even though there is excess significance for Y causes X, the meta-regression models do not lead to false-positive findings of genuine Granger causality. Compared to the previous case without p-hacking, the type I errors of the basic model are even smaller indicating the increased presence of overfitted VAR models that increase the downward bias of β_B^{gc}. But the type I errors of the extended model are increased so that there is now a greater difference between the basic and extended models. The type I errors of both meta-regression models show little reaction to the degree of p-hacking except when h = 100, and even then the errors are smaller not larger. DGP1 shows the same patterns as DGP2 but with generally lower power and a smaller difference between the two meta-regression models (see Appendix A2).

Figure 9 shows how the power of both models varies with the degree of p-hacking for DGP2. p-hacking based on exploiting sampling errors and overfitting lag lengths has little impact on the power of both meta-regression models. They reliably identify whether statistically significant Granger causality tests are based on genuine Granger causality or based on p-hacking. DGP1 shows the same patterns as DGP2 (see Appendix A2).

The simulation findings are similar to those of DGP1a and, therefore, are not reported. Notable differences are that a lag length of 1 has the highest frequency to occur for both AIC and BIC, correlated errors increase the difference between the basic and extended meta-regression model, and type I errors have a tendency to be larger and to exceed 0.05 if BIC is used in the primary studies.

[8] We also analyzed a case in which primary studies select for any statistically significant Granger causality test irrespective of the direction of causality. In this case almost no selection bias occurs as genuine Granger causality is present in all DGPs and this genuine Granger causality usually provides a statistically significant Granger causality test that can be selected for publication.

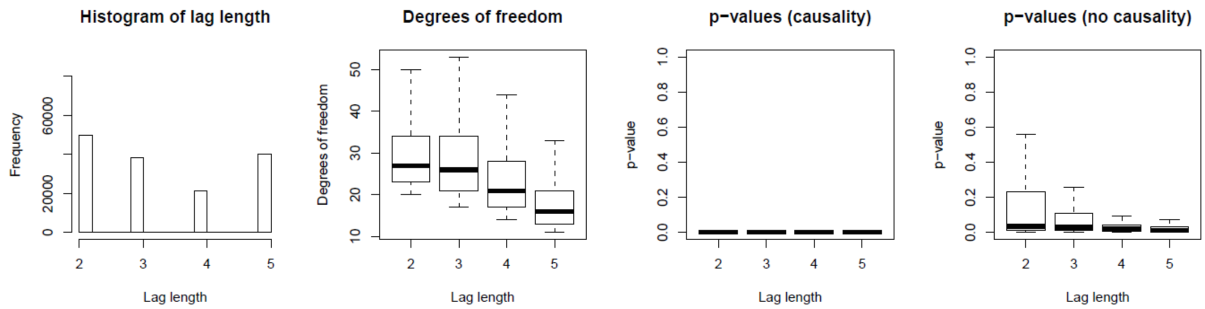

Fig. 7 Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP2a in the presence of theory-confirmation bias ($h = 75$). See caption of Fig. 4 for further details.

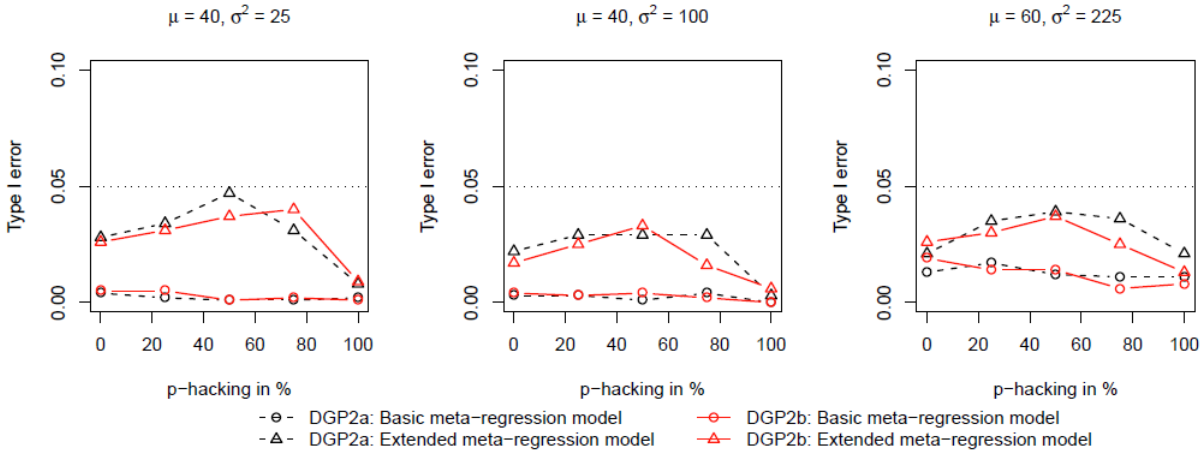

Fig. 8 Type I errors of both the basic and extended meta-regression models for DGP2a and DGP2b in the presence of theory-confirmation bias are shown. Type I errors of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = \mathrm{I}$ are reported in relation to the share of *p*-hacked studies ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample size distributions in column one and two and a larger primary sample size distribution in column three.

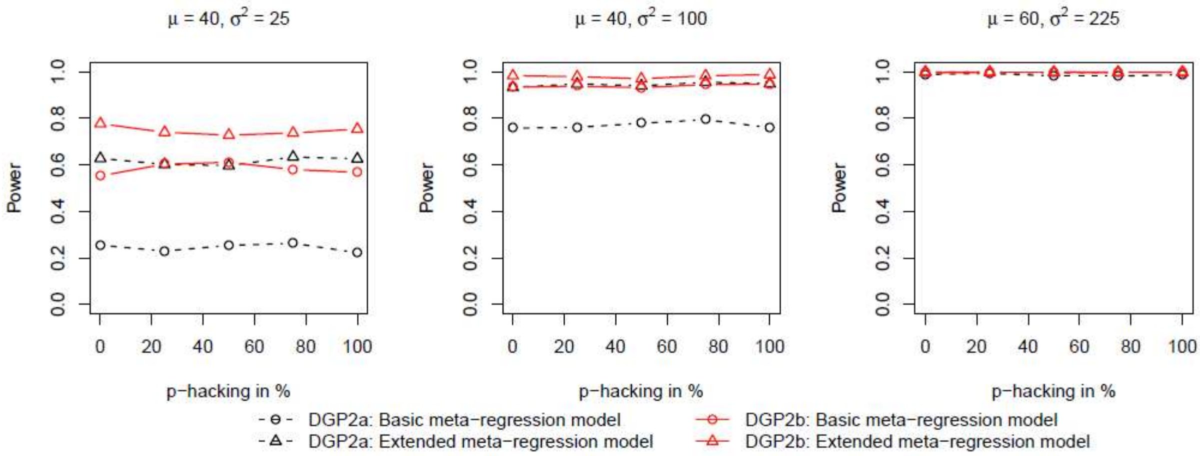

Fig. 9 Power of both the basic and extended meta-regression model for DGP2 in the presence of theory-confirmation bias is shown. Power curves of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black)

and DGP2b (red) with $\Omega = \mathrm{I}$ are reported in relation to the share of p-hacked studies ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample size distributions in column one and two and a larger primary sample size distribution in column three.

4. p-hacking in the energy-growth literature

4.1. Background and data

In this section, we investigate the source of statistically significant Granger causality tests in the literature that explores the relationship between energy use and economic output. We select studies that use the Toda-Yamamoto procedure from the data set compiled by Bruns *et al.* (2014). Appendix A3 provides an overview of these 23 studies. As many studies report multiple estimates, the data set contains 126 Granger causality statistics in each direction. There are 66 test statistics based on a lag length of one, 26 based on a lag length of two, and 34 that use a lag length of three for each direction of causality.[9]

The average z^{gc} value in the sample for energy causes growth tests is 0.83, which corresponds to an average p-value of 0.20. The average z^{gc} value for growth causes energy tests is 1.03, which corresponds to an average p-value of 0.15. Both average p-values are considerably lower than we would expect in the absence of genuine Granger causality (average p-value = 0.5). Can this high level of average significance be explained by the presence of genuine Granger causality?

We group the test statistics into three categories according to the primary VAR specifications used (Table 2). We have 66 observations that use a bivariate specification with energy consumption and economic output only. 19.70% of these bivariate specifications are statistically significant at the 0.05 level for a test of energy causes growth and 27.27% for a test of growth causes energy. The degrees of freedom are reasonably large and the chosen lag length small. We have 41 observations that use a primary VAR specification with capital and labor as additional control variables. In each direction of causality, almost half of these statistics are statistically significant at the 0.05 level. In addition, compared to the bivariate specification the number of degrees of freedom is low and the lag lengths are high. Finally, we have a third category that contains all remaining primary VAR specifications with various

[9] We delete two test statistics from Esso (2010) as they are the only tests using a VAR model with a lag length of four in our sample.

control variables (CO_2 emissions, energy prices, labor, capital, and population) but insufficient observations to group them into separate categories.

Table 2 Properties of Granger causality tests

Control variables	Number of tests	Number of studies	Energy causes growth		Growth causes energy		Percentiles of df			Number of lags		
			$p < 0.05$	$p < 0.1$	$p < 0.05$	$p < 0.1$	25	50	75	1	2	3
None	66	6	0.20	0.23	0.27	0.38	28	35	38	47	18	1
Capital and Labor	41	7	0.49	0.51	0.46	0.56	12	14	21	7	5	29
Other	19	10	0.11	0.21	0.37	0.42	17	21	28.5	12	3	4

Notes: df denotes degrees of freedom and p denotes p-value

4.2. Meta-regression analysis

Granger causality tests are sensitive to the set of other relevant information taken into account (Granger, 1988). If researchers omit relevant variables they may obtain spurious findings of causality (Lütkepohl 1982; Stern, 1993). In the presence of omitted-variable biases in the primary literature, meta-regression models will also detect spurious "genuine effects" (Bruns, 2017). By controlling for the different VAR specifications used in the primary literature the meta-analyst can use the meta-regression model to investigate whether a positive relation between z^{gc} and $\sqrt{df}$ is due to omitted-variable bias or a genuine effect.[10]

Furthermore, the addition of control variables to the primary VAR specification can deplete the degrees of freedom increasing the probability of overfitting the VAR model and obtaining spuriously significant Granger causality tests. In general, adding variables to the VAR model increases the penalty terms of the information criteria. But if the addition of variables is used to deplete the *df* leading to very low *df*, the increased variance of $ln|\hat{\Sigma}_{p^*}| - ln|\hat{\Sigma}_{p^*+h}|$ in (3) may exceed the increase in the penalty term implying a higher probability of overfitting

We generalize the extended meta-regression model (6) to take the dependence between the Granger causality test statistics and the three primary VAR specifications into account, using the following regression:

[10] If some relevant variables are not included by any primary study, it is impossible to identify a genuine effect using meta-regression analysis. Instead, meta-regression analysis may indicate the need for further research.

$$z_i^{gc} = \alpha_1^{gc} + \beta_1^{gc}\sqrt{df_i} + D_{KL}\left(\alpha_2^{gc} + \beta_2^{gc}\sqrt{df_i}\right) + D_{Ot}\left(\alpha_3^{gc} + \beta_3^{gc}\sqrt{df_i}\right) + \gamma^{gc} p_i + \varepsilon_i^{gc} \quad (7)$$

where $D_{KL} = 1$ if capital and labor are used as control variables in the primary VAR specification and is zero otherwise and $D_{Ot} = 1$ if control variables other than capital and labor are used and is zero otherwise.[11] We control for the lag lengths of the underlying VAR models with one continuous variable as the Granger causality test by Toda and Yamamoto (1995) results in over-rejection (under-rejection) of the null of Granger non-causality if the lag length is overfitted (underfitted) for both the presence and absence of genuine Granger causality. Accordingly, $H_0\colon \beta_1^{gc} \leq 0$ tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ if the bivariate VAR specification was used and $H_0\colon \beta_1^{gc} + \beta_2^{gc} \leq 0$ tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ if capital and labor are used as control variables.

We carry out the inferences by using confidence intervals. The aim is to shift attention from statistical significance to the size of the coefficients (Cumming, 2014). Moreover, we bootstrap these confidence intervals, as the results of our Monte Carlo simulations (Section 3) indicate that both the basic and extended meta-regression model are under-sized, i.e. they reject the null less than the nominal significance level. Bootstrapping is known to perform well in these situations (MacKinnon, 2002).[12] We use the bias-corrected and accelerated (BC_a) bootstrap to construct confidence intervals for each coefficient using 1000 iterations (Efron, 1978; DiCiccio and Efron, 1996).

4.3. Results

Table 3 presents the results of the meta-regression models for energy causes growth and *vice versa*.[13] Columns (1) present the basic model. Here, the estimate of β_B^{gc} is negative and the estimate of the constant is positive, as we would expect in the presence of overfitted lag lengths and the corresponding over-rejection for small df. The test for a positive relation between z_i^{gc}

[11] Ideally, we would control for every different combination of primary control variables used in the literature. Unfortunately, the number of observations for most of these is very small. For example, only one article in our sample of Toda-Yamamoto tests controls for energy prices. Therefore, we have lumped primary studies with various control variables together into another category.

[12] We are thankful to an anonymous reviewer for making this point.

[13] We also conducted the analysis excluding Vaona (2010) who has the largest values of df – 127 and 130 –more than double the next highest value of 49. The results remain qualitatively the same and are reported in Appendix A4. They indicate a stronger influence of overfitted lag lengths on the inference of the meta-regression models as we would expect when dropping observations with large df.

and $\sqrt{df_i}$ is one-sided but the reported confidence intervals do not show evidence for such a positive relation for energy causes growth and *vice versa*. Columns (2) show the extended model. Adding the lag length as a control variable leads to estimates of both β_E^{gc} and the constant that are close to 0 with the 0.95 confidence intervals including 0. As expected, the coefficient of the lag length variable is positive, and the 0.95 confidence interval does not include 0. Columns (3) show the generalized extended model (7) that tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ for each of the three primary VAR specification categories. For both energy causes growth and *vice versa*, we calculated the 0.90 confidence intervals of β_1^{gc} and $\beta_1^{gc} + \beta_2^{gc}$ as the lower bound of these confidence intervals correspond to a one sided t-test of a positive relation between z_i^{gc} and $\sqrt{df_i}$ at the 0.05 significance level for the specification without control variables and for the specification with capital and labor as control variables. For energy causes growth, these confidence intervals are [-0.33,0.14] and [-0.46, 0.41] and for growth causes energy, [-0.43, 0.30] and [-0.43, 0.34] indicating no evidence for a positive relation between z_i^{gc} and $\sqrt{df_i}$.

Fig. 10 shows that a lag length of three predominantly occurs for small *df*, and Granger causality tests obtained by a VAR model with a lag length of three tend to result in larger values of z^{gc}. As outlined in Table 2, a lag length of three occurs almost exclusively for the primary VAR specification with capital and labor and the Granger causality tests with these control variables also have the highest levels of statistical significance, whereas Granger causality tests for VARs with capital and labor but smaller lag lengths tend to be non-significant. This indicates that additional control variables might be used to deplete *df* resulting in overfitted VAR models with statistically significant Granger causality tests. Given that the probability of overfitting increases with decreasing *df*, adding control variables to the primary VAR specification may facilitate the search for statistically significant results.

This empirical application shows that there is no evidence for a genuine relation between energy use and economic output in bivariate VAR specifications or in VAR specifications with capital and labor as control variables – at least in this linear setup. But we find evidence that overfitted lag lengths and the corresponding false-positive findings of Granger causality are present in this literature.

Both VAR specifications (bivariate and with capital and labor) may suffer from omitted-variable biases that obscure a genuine relation. Bruns *et al*. (2014) find some evidence that there appears to be genuine Granger causality from economic output to energy use if energy prices are controlled for, which mimics an energy demand function. Further research is needed to validate this finding.

Table 3 Results of meta-regression models

	Energy causes growth			Growth causes energy		
	(1)	(2)	(3)	(1)	(2)	(3)
Constant	**2.32**	-0.16	0.04	**2.09**	-0.55	-0.35
	(1.10, 3.46)	(-1.64, 1.48)	(-1.44, 2.23)	(0.60, 3.62)	(-2.22, 2.01)	(-2.83, 2.54)
df	**-0.28**	-0.05	-0.06	-0.20	0.04	0.02
	(-0.49, -0.07)	(-0.27, 0.16)	(-0.42, 0.16)	(-0.48, 0.07)	(-0.30, 0.29)	(-0.46, 0.39)
Lags		**0.73**	**0.52**		**0.77**	**0.70**
		(0.36, 1.06)	(0.11, 1.04)		(0.32, 1.20)	(0.19, 1.21)
KL			0.47			0.44
			(-3.03, 3.51)			(-2.99, 3.89)
KL**df*			0.04			-0.07
			(-0.50, 0.66)			(-0.65, 0.50)
Other			-1.18			-2.97
			(-5.49, 4.62)			(-8.76, 2.12)
Other**df*			0.22			0.63
			(-0.90, 1.02)			(-0.42, 1.79)
Obs.	126	126	126	126	126	126
Adj. R^2	0.06	0.17	0.18	0.02	0.13	0.12

Notes: Bootstrapped 0.95 confidence intervals in parentheses. Coefficients whose confidence intervals do not include 0 are in bold.

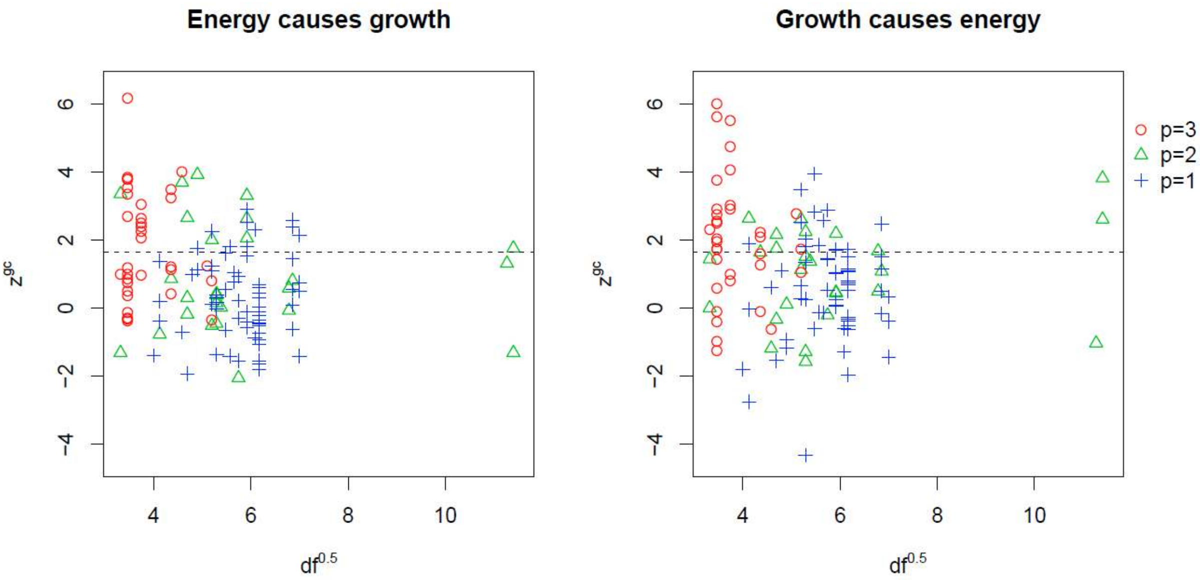

Fig. 10 Relations of lag lengths, degrees of freedom, and levels of statistical significance in the empirical meta-sample are shown. The z^{gc} values are reported as function of $\sqrt{df}$ for a lag length of one ($p = 1$), two ($p = 2$), and three ($p = 3$). The dashed line is at 1.64 separating the graph into statistically significant Granger causality tests (above) and statistically non-significant Granger causality tests (below) at the 0.05 level of significance.

5. Discussion

We show that overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality compared to a VAR with the correct lag length occur frequently in small to moderate sample sizes. This hampers inference on the presence of genuine Granger causality using meta-regression models. We show that the extended meta-regression model can adjust for overfitted lag lengths and improves power compared to the basic meta-regression model.

The simulation results reveal that the basic meta-regression model finds it difficult to detect small genuine causal coefficients, as these are interpreted as the absence of genuine Granger causality. The extended model provides an improvement in power particularly for small genuine effects as it takes the overfitting into account. Economic effects are often small, highlighting how important the correction for overfitting is. Our application indicates no evidence for a genuine effect. These findings are supported by Bruns *et al*. (2014) who included "the degrees of freedom lost in fitting the model" as a control variable in their meta-regression model so that the square root of the degrees of freedom variable only reflects variation in the degrees of freedom due to variation in the sample size. This control variable is mainly determined by the chosen lag length and by the number of control variables added to the VAR model. The approach discussed here allows us to further disentangle the sources of spuriously statistically significant Granger causality tests. Overfitting of the lag length can occur in bivariate VAR specifications with small sample sizes where the degrees of freedom lost in fitting the model may be low. Conversely, it is unlikely that overfitted lag lengths occur even if the degrees of freedom lost in fitting the model are large when the sample size is also large. In practice, the approach of Bruns *et al*. (2014) may or may not correlate with the approach used here depending on the sample. For our sample, the correlation coefficient between the number of lags and the degrees of freedom lost in fitting the model is 0.89 but the correlation need not be this high, particularly for higher-dimensional VAR models.

As demonstrated here, meta-regression models can be a powerful tool for detecting biases and identifying genuine empirical effects. But challenges remain in the application of meta-regression models to observational data (Bruns, 2017). More research is needed to better understand how meta-regression models can deal with various sets of control variables in the primary studies. We tested for a positive relation between probit transformed p-values and the square root of the degrees of freedom for each set of primary control variables by using dummy variables. If we would have found such a positive relationship, we could have then discussed whether the set of control variables is adequate or whether omitted-variable biases are likely to have caused this positive relationship. But this approach is only feasible if multiple primary studies with the same control variables are present in the literature, which may often not be the case, as publication requires novelty, which often means the inclusion of different control variables.

The application reveals that researchers may add control variables to the VAR model to deplete the degrees of freedom resulting in an increased probability of generating false positive findings of Granger causality. While overfitted lag lengths can be used to p-hack, VAR models with overfitted lag lengths are not necessarily the result of p-hacking but they also occur in the use of standard research designs outlined in textbooks. For example, capital and labor are reasonable control variables to include in a test of Granger causality between energy use and economic output and the false-positive findings of Granger causality for this specification may be the result of researchers trying to estimate a better model.

Generally, our findings contribute to the increasing body of evidence that biases and p-hacking may be prevalent in empirical economics research. They indicate the need for measures that improve the reliability and credibility of empirical research. One of these measures is to de-emphasize null-hypothesis significance testing (for an overview see Cumming, 2014). As prominently pointed out by McCloskey and Ziliak (1996), statistical significance is often falsely considered to represent economic significance. Researchers tend to chase *p*-values that are below some common threshold of statistical significance while the economic interpretability of the effect size remains neglected. Their critique is relevant to literatures that focus on Granger causality tests and largely ignore effect sizes. This is even more important where sample sizes are very large compared to macroeconomics, such as is often the case in finance or neuroscience. Even if a genuine effect is considered to be absent, (very) large sample sizes may often generate statistically significant estimates as even tiny biases will generate arbitrarily small *p*-values (Schuemie *et al.*, 2014; Kim and Ji, 2015; Bruns and Ioannidis, 2016).

6. Conclusions

By modeling the complete process of Granger causality testing, we show that overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality are prevalent in a variety of scenarios mirroring research in macroeconomic time series analysis. Overfitting leaves empirical researchers with uncertainty about the reliability of inferences. Particularly, *p*-hacking based on overfitted lag lengths can lead to excess significance even though genuine Granger causality is absent. We introduce a meta-regression model that controls for spurious significance generated by overfitted lag lengths. The suggested model has higher power than the basic meta-regression model and both provide adequate type I errors.

We apply the suggested meta-regression model to the large literature that tests for Granger causality between energy consumption and economic output. We generalize the meta-regression models to the synthesis of different multivariate VAR models and find that this empirical literature shows no evidence for genuine Granger causality even though excess significance is present. Specifically, we find evidence that adding control variables to the primary VAR models can be used to deplete the degrees of freedom, which increases the probability of obtaining false-positive findings of Granger causality due to overfitted lag lengths.

Data and computer code availability

The data and code used in this paper (1. Code, 2. Data, 3. Detailed readme files) are collected in the electronic supplementary material of this article.

References

Adom, P. K. (2011). Electricity consumption-economic growth nexus: The Ghanaian case. International Journal of Energy Economics and Policy, 1(1):18-31.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716-723.

Alam, M., Begum, I., Buysse, J., Rahman, S., and Van Huylenbroeck, G. (2011). Dynamic modeling of causal relationship between energy consumption, CO_2 emissions and economic growth in India. Renewable and Sustainable Energy Reviews 15(6):3243-3251.

Bowden, N. and Payne, J. (2009). The causal relationship between US energy consumption and real output: A disaggregated analysis. Journal of Policy Modeling, 31(2):180-188.

Brodeur, A., Le, M., Sangnier, M. and Zylberberg, Y. (2016). Star wars: The empirics strike back, American Economic Journal: Applied Economics, 8(1):1-32.

Bruns, S. B. (2017). Meta-regression models and observational research. Oxford Bulletin of Economics and Statistics.

Bruns, S. B. and Ioannidis, J. P. A. (2016). p-Curve and p-Hacking in Observational Research, PLoS ONE 11:e0149144.

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.

Ciarreta, A., Otaduy, J. and Zarraga, A. (2009). Causal relationship between electricity consumption and GDP in Portugal: a multivariate approach. Empirical Economics Letters, 8(7):693-701.

Cumming, G. (2014). The New Statistics: Why and How, Psychological Science, 25(1): 7-29.

DiCiccio, T. J. and B. Efron (1996). Bootstrap Confidence Intervals, Statistical Science, 11(3): 189-212.

Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82(397): 171-185.

Esso, L. J. (2010). Threshold cointegration and causality relationship between energy use and growth in seven African Countries. Energy Economics, 32(6):1383-1391.

Glaeser, E. L. (2011). 'Researcher incentives and empirical methods', In: Schotter, A. and Caplin, A. (Eds.), The foundations of positive and normative economics: A hand book. Oxford University Press.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424-438.

Granger, C. W. J. (1988) Some recent developments in a concept of causality. Journal of Econometrics 39:199-211.

Hacker, R. S. and Hatemi-J, A. (2008). Optimal lag-length choice in stable and unstable VAR models under situations of homoscedasticity and ARCH. Journal of Applied Statistics, 35(6):601-615.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika, 76:297–307.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine 2(8):e124.

Ioannidis, J. P. A. and Doucouliagos, C. (2013). What's to know about credibility of empirical economics? Journal of Economic Surveys, 27(5): 997–1004.

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, C. (2016). The power of bias in economics research. Economics Series, SWP 2016/1.

Kim, J. H. and Ji, P. (2015). Significance testing in empirical finance: a critical review and assessment. Journal of Empirical Finance, 34: 1-14.

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

Lee, C. (2006). The Causality relationship between energy consumption and GDP in G-11 countries revisited. Energy Policy, 34(9):1086-1093.

Lotfalipour, M., Falahi, M. and Ashena, M. (2010). Economic growth, co_2 emissions, and fossil fuels consumption in Iran. Energy, 35(12):5115-5120.

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

Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. Journal of Time Series Analysis, 6(1):35-52.

Lütkepohl, H. (2007). New introduction to multiple time series analysis. Springer.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. Canadian Journal of Economics, 35(4): 615-644.

McCloskey, D. and Ziliak, S. (1996), The standard error of regressions Journal of Economic Literature, 34: 97–114.

Mehrara, M. (2007). Energy consumption and economic growth: The case of oil exporting countries. Energy Policy, 35(5):2939-2945.

Menyah, K. and Wolde-Rufael, Y. (2010a). CO_2 emissions, nuclear energy, renewable energy and economic growth in the US. Energy Policy, 38(6):2911-2915.

Menyah, K. and Wolde-Rufael, Y. (2010b). Energy consumption, pollutant emissions and economic growth in South Africa. Energy Economics, 32(6):1374-1382.

Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: some evidence and implications. Journal of Monetary Economics, 10(2), 139-162.

Nickelsburg, G. (1985). Small-sample properties of dimensionality statistics for fitting VAR models to aggregate economic data: A Monte Carlo study. Journal of Econometrics, 28(2):183-192.

Ozcicek, O. and Mcmillin, W. (1999). Lag length selection in vector autoregressive models: symmetric and asymmetric lags. Applied Economics, 31(4):517-524.

Payne, J. E. (2009). On the dynamics of energy consumption and output in the US. Applied Energy, 86(4):575-577.

Payne, J. E. and Taylor, J. P. (2010). Nuclear energy consumption and economic growth in the US: an empirical note. Energy Sources, Part B: Economics, Planning, and Policy, 5(3):301-307.

Sari, R. and Soytas, U. (2009). Are global warming and economic growth compatible? evidence from five OPEC countries. Applied Energy, 86(10):1887-1893.

Schuemie, M. J., Ryan, P. B., Dumouchel, W., Suchard, M. A., and Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. Statistics in Medicine, 33(2):209-218.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461-464.

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

Soytas, U. and Sari, R. (2009). Energy consumption, economic growth, and carbon emissions: Challenges faced by an EU candidate member. Ecological Economics, 68(6):1667-1675.

Soytas, U., Sari, R. and Ewing, B. (2007). Energy consumption, income, and carbon emissions in the United States. Ecological Economics, 62(3-4):482-489.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxford Bulletin of Economics and Statistics, 70(1):103-127.

Stanley, T. D. and Jarrell, S. B. (1989). Meta-regression analysis: A quantitative method of literature surveys. Journal of Economic Surveys, 3(2):161-170.

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

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

Vaona, A. (2012). Granger Non-causality between (non)renewable energy consumption and output in Italy since 1861: The (ir)relevance of structural breaks. Energy Policy, 45:226-236.

Vivalt, E. (2017). The trajectory of specification searching and publication bias across methods and disciplines. Working paper.

Wasserstein R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. The American Statistician, 70(2): 129-133.

Wolde-Rufael, Y. (2009). Energy consumption and economic growth: the experience of African countries revisited. Energy Economics, 31(2):217-224.

Wolde-Rufael, Y. (2010a). Bounds test approach to cointegration and causality between nuclear energy consumption and economic growth in India. Energy Policy, 38(1):52-58.

Wolde-Rufael, Y. (2010b). Coal consumption and economic growth revisited. Applied Energy, 87(1):160-167.

Wolde-Rufael, Y. and Menyah, K. (2010). Nuclear energy consumption and economic growth in nine developed countries. Energy Economics, 32(3):550-556.

Zachariadis, T. (2007). Exploring the relationship between energy use and economic growth with bivariate models: New Evidence from G-7 Countries. Energy Economics, 29(6):1233-1253.

Zhang, X.-P. and Cheng, X.-M. (2009). Energy consumption, carbon emissions, and economic growth in China. Ecological Economics, 68(10):2706-2712.

Zapata, H. O. and Rambaldi, A. N. (1997). Monte Carlo evidence on cointegration and causation. Oxford Bulletin of Economics and Statistics, 59(2):285-298.

Ziramba, E. (2009). Disaggregate energy consumption and industrial production in South Africa. Energy Policy, 37(6):2214-2220.

Appendix A1

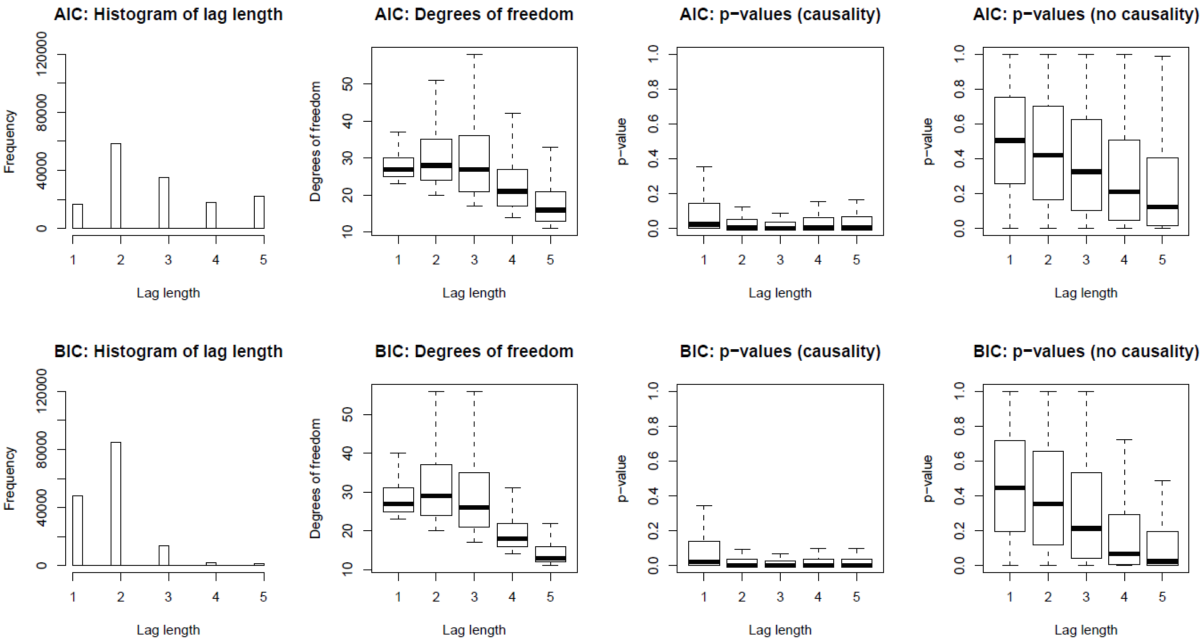

Fig. 4a Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP1a. The first column shows the histograms of selected lag lengths in simulated primary studies across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The box represents the interquartile range and the whiskers extend to the largest data point within 1.5 times the interquartile range. The third column shows the boxplots of p-values in simulated primary studies for the presence of Granger causality, whereas the fourth column presents the boxplots of p-values in the absence of Granger causality.

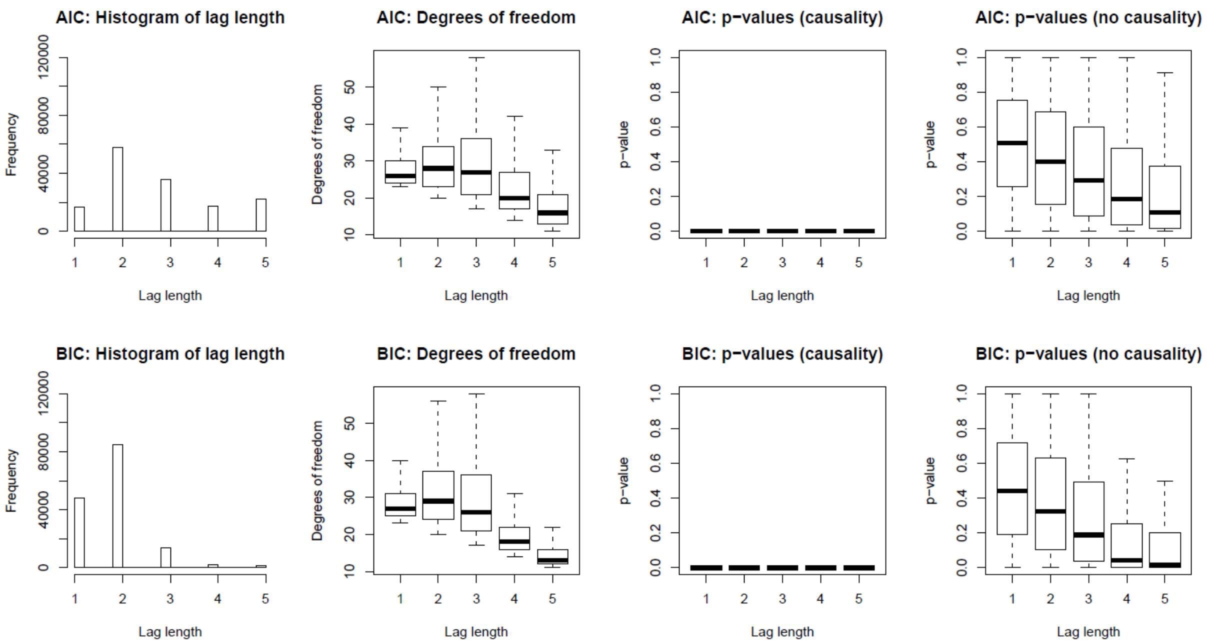

Fig. 4b Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP1b. The first column shows the histograms of selected lag lengths in simulated primary studies across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The box represents the interquartile range and the whiskers extend to the largest data point within 1.5 times the interquartile range. The third column shows the boxplots of p-values in simulated primary studies for the presence of Granger causality, whereas the fourth column presents the boxplots of p-values in the absence of Granger causality.

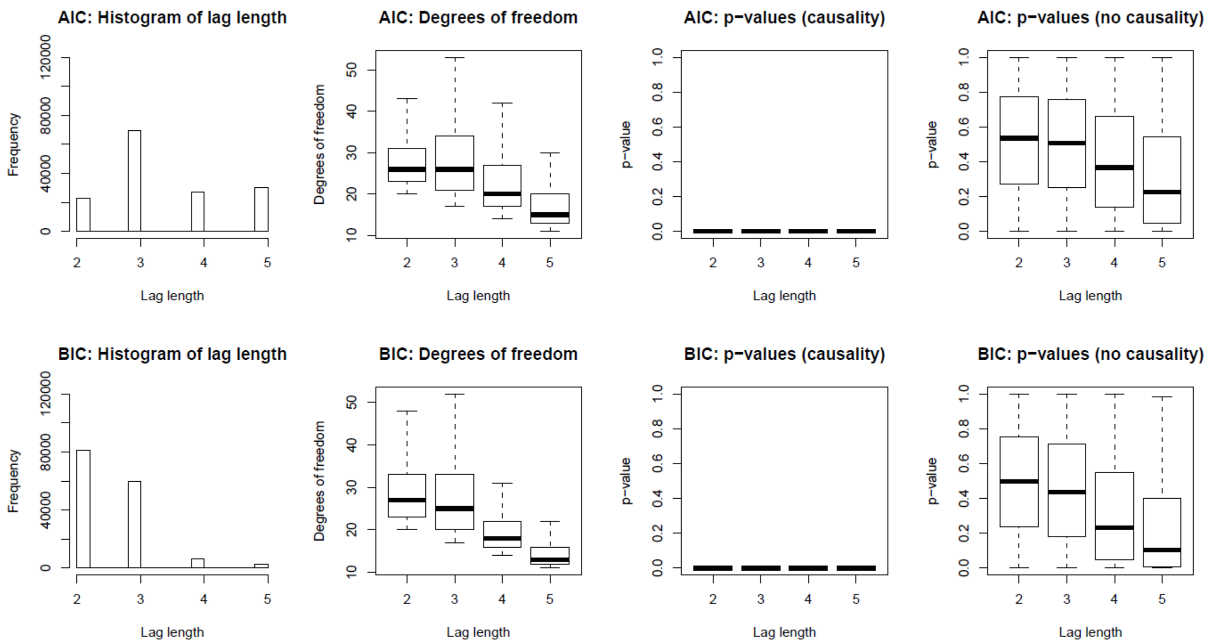

Fig. 4c Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP2b. The first column shows the histograms of selected lag lengths in simulated primary studies across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The box represents the interquartile range and the whiskers extend to the largest data point within 1.5 times the interquartile range. The third column shows the boxplots of p-values in simulated primary studies for the presence of Granger causality, whereas the fourth column presents the boxplots of p-values in the absence of Granger causality. A lag length of one was selected for less than 0.1% of primary studies and these findings are not reported.

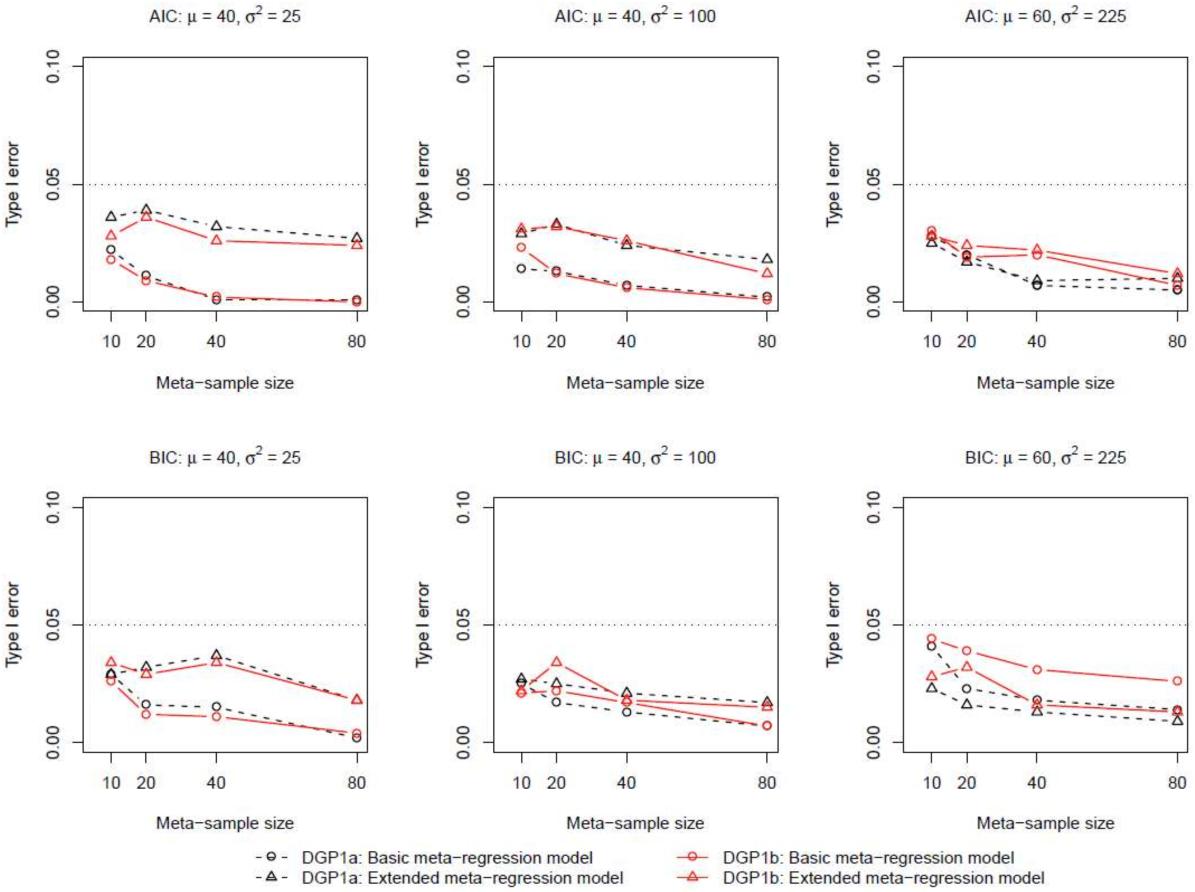

Fig. 5a Type I errors of both the basic and extended meta-regression models for DGP1a and DGP1b are shown. Type I errors of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega =$ I are reported if the AIC (upper row) or the BIC (lower row) is used for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Figure 6a: Power of Meta-Regression Models for DGP1a and DGP1b

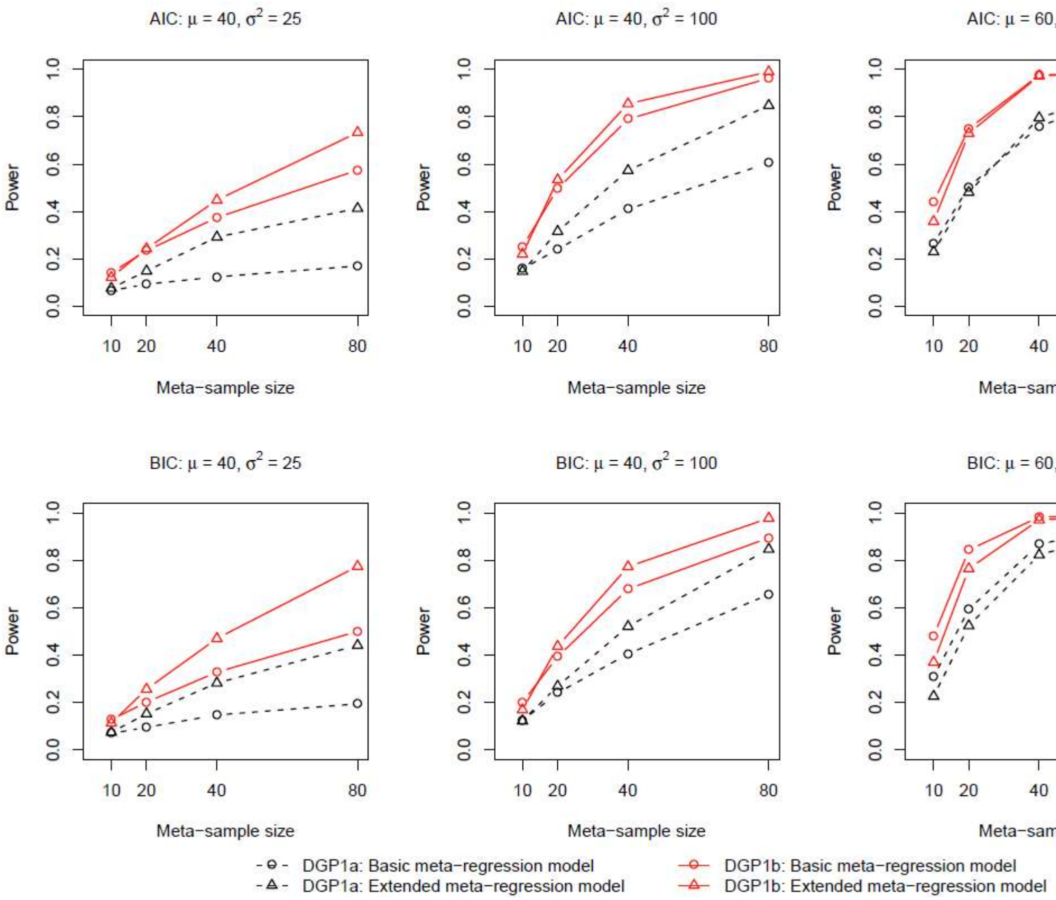

Fig. 6a Power of both the basic and extended meta-regression models for DGP1a and DGP1b are shown. Power curves of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega = \mathrm{I}$ are reported if the AIC (upper row) or the BIC (lower row) is used for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Appendix A2

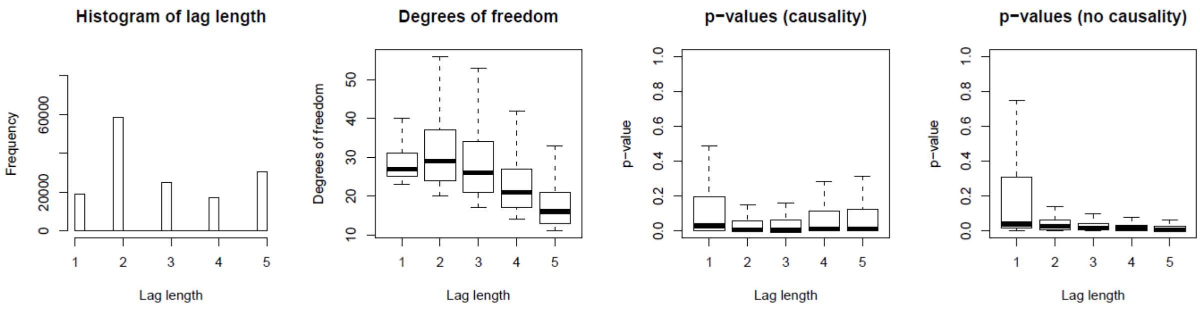

Fig. 7a Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP1a in the presence of theory-confirmation bias ($h = 75$). See caption of Fig. 4 for further details.

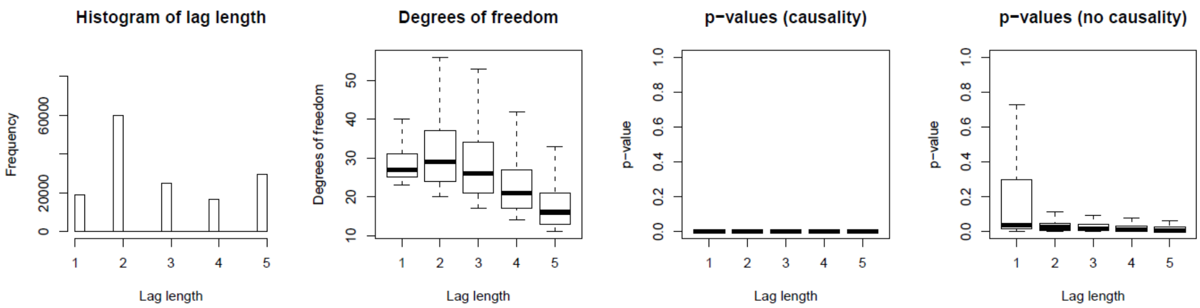

Fig. 7b Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-causality is shown for DGP1b in the presence of theory-confirmation bias ($h = 75$). See caption of Fig. 4 for further details.

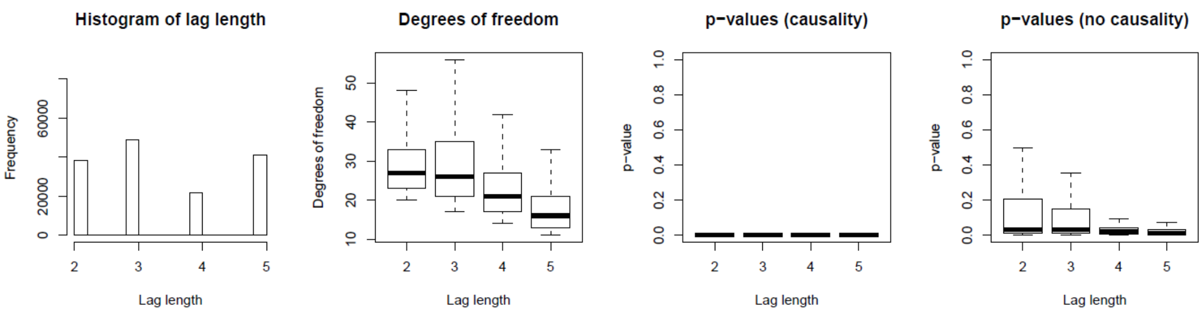

Fig. 7a Prevalence of overfitted lag lengths and the corresponding over-rejection of the null of Granger non-

causality is shown for DGP2b in the presence of theory-confirmation bias ($h = 75$). See caption of Fig. 4 for further details.

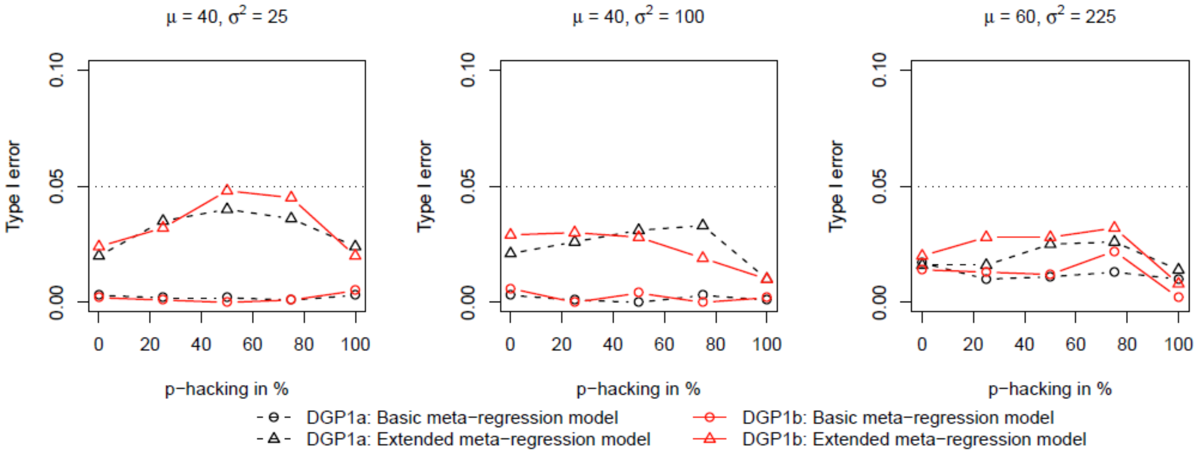

Fig. 8a Type I errors of both the basic and extended meta-regression models for DGP1a and DGP1b in the presence of theory-confirmation bias are shown. Type I errors of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega = \mathrm{I}$ are reported in relation to the share of p-hacked studies ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample size distributions in column one and two and a larger primary sample size distribution in column three.

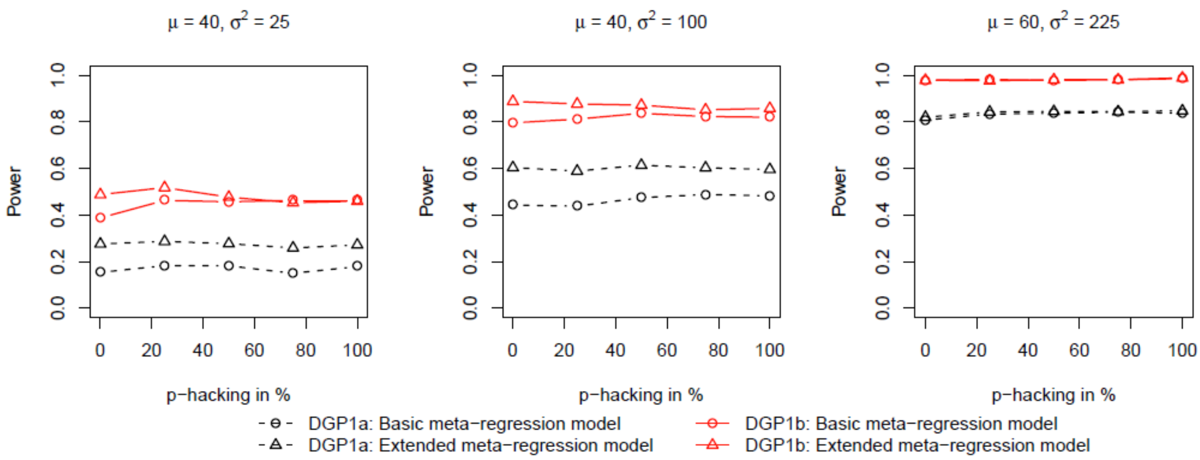

Fig. 9a Power of both the basic and extended meta-regression model for DGP1a and DGP1b in the presence of theory-confirmation bias is shown. Power curves of $H_0\colon \beta_B^{gc} \leq 0$ (circles) and $H_0\colon \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega = \mathrm{I}$ are reported in relation to the share of p-hacked studies ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample size distributions in column one and two and a larger primary sample size distribution in column three.

1

Appendix A3

Table 4: Studies included in the empirical application

Authors and date	Countries	Control variables
Adom (2011)	GHA	-
Alam *et al*. (2011)	IND	Employment, capital, CO_2
Bowden and Payne (2009)	USA	Employment, capital
Ciarreta *et al*. (2009)	PRT	Energy price
Esso (2010)	CMR; COG; CIV; GHA; KEN; ZAF	-
Lee (2006)	G-11 countries	-
Lotfalipour *et al*. (2010)	IRN	CO_2
Mehrara (2007)	IRN, KWT, SAU	-
Menyah and Wolde-Rufael (2010a)	USA	CO_2
Menyah and Wolde-Rufael (2010b)	ZAF	Capital, CO_2
Payne (2009)	USA	Employment, capital
Payne (2010)	USA	Employment, capital
Sari and Soytas (2009)	DZA, IND, NGA, SAU, VEN	Employment, CO_2
Soytas *et al*. (2007)	USA	Employment, capital, CO_2
Soytas and Sari (2009)	TUR	Employment, capital, CO_2
Vaona (2012)	ITA	-
Wolde-Rufael (2009)	17 African countries	Employment; capital
Wolde-Rufael (2010a)	IND	Employment; capital
Wolde-Rufael (2010b)	CHN; IND; JPN; KOR; ZAF; USA	Employment; capital
Wolde-Rufael and Menyah (2010)	9 developed countries	Employment; capital
Zachariadis (2007)	G7 countries	-
Zhang and Cheng (2009)	CHN	Capital; CO_2; population
Ziramba (2009)	ZAF	Employment

Table 5: Results of the meta-regression models without Vaona (2010)

	Energy causes Growth			Growth causes Energy		
	(1)	(2)	(3)	(1)	(2)	(3)
Constant	**3.20**	-0.39	0.80	**3.30**	0.08	2.42
	(1.84, 4.71)	(-2.73, 1.91)	(-2.67, 3.84)	(1.86, 4.85)	(-2.76, 2.53)	(-2.42, 6.62)
Df	**-0.46**	-0.02	-0.18	**-0.44**	-0.05	-0.43
	(-0.72, -0.21)	(-0.35, 0.35)	(-0.64, 0.34)	(-0.72, -0.20)	(-0.38, 0.34)	(-1.07, 0.33)
lags		**0.76**	0.48		**0.68**	**0.57**
		(0.39, 1.21)	(-0.004, 1.02)		(0.21, 1.12)	(0.01, 1.12)
KL			-0.11			-1.68
			(-3.68, 3.33)			(-5.29, 3.03)
KL**df*			0.14			0.31
			(-0.47, 0.79)			(-0.45, 0.93)
Other			-1.84			-5.37
			(-6.89, 4.22)			(-11.63, 0.78)
Other**df*			0.33			1.04
			(-0.94, 1.25)			(-0.12, 2.35)
Obs.	123	123	123	123	123	123
Adj. R^2	0.10	0.17	0.18	0.08	0.13	0.13

Notes: Bootstrapped 0.95 confidence intervals in parentheses. Coefficients whose confidence intervals do not include 0 are in bold.