Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Meta-Granger Causality Testing

CAMA Working Paper 22/2015
June 2015

Stephan B. Bruns
University of Kassel, Germany

David I. Stern
Crawford School of Public Policy, ANU
Centre for Applied Macroeconomic Analysis (CAMA), ANU

Abstract

Understanding the (causal) mechanisms at work is important for formulating evidence-based policy. But evidence from observational studies is often inconclusive with many studies finding conflicting results. In small to moderately sized samples, the outcome of Granger causality testing heavily depends on the lag length chosen for the underlying vector autoregressive (VAR) model. Using the Akaike Information Criterion, there is a tendency to overfit the VAR model and these overfitted models show an increased rate of false-positive findings of Granger causality, leaving empirical economists with substantial uncertainty about the validity of inferences. We propose a meta-regression model that explicitly controls for this overfitting bias and we show by means of simulations that, even if the primary literature is dominated by false-positive findings of Granger causality, the meta-regression model correctly identifies the absence of genuine Granger causality. We apply the suggested model to the large literature that tests for Granger causality between energy consumption and economic output. We do not find evidence for a genuine relation in the selected sample, although excess significance is present. Instead, we find evidence that this excess significance is explained by overfitting bias.

Keywords

Granger causality, vector autoregression, information criteria, meta-analysis, meta-regression, bias, publication selection bias

JEL Classification

Address for correspondence:

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

Meta-Granger Causality Testing

Stephan B. Bruns

Meta-Research in Economics Group, University of Kassel, Nora-Platiel-Str. 5, 34109 Kassel, Germany. E-mail: bruns@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. E-mail: david.stern@anu.edu.au. Phone: +61-2-6125-0176.

Abstract

Understanding the (causal) mechanisms at work is important for formulating evidence-based policy. But evidence from observational studies is often inconclusive with many studies finding conflicting results. In small to moderately sized samples, the outcome of Granger causality testing heavily depends on the lag length chosen for the underlying vector autoregressive (VAR) model. Using the Akaike Information Criterion, there is a tendency to overfit the VAR model and these overfitted models show an increased rate of false-positive findings of Granger causality, leaving empirical economists with substantial uncertainty about the validity of inferences. We propose a meta-regression model that explicitly controls for this overfitting bias and we show by means of simulations that, even if the primary literature is dominated by false-positive findings of Granger causality, the meta-regression model correctly identifies the absence of genuine Granger causality. We apply the suggested model to the large literature that tests for Granger causality between energy consumption and economic output. We do not find evidence for a genuine relation in the selected sample, although excess significance is present. Instead, we find evidence that this excess significance is explained by overfitting bias.

Keywords: Granger causality, vector autoregression, information criteria, meta-analysis, meta-regression, bias, publication selection bias

Acknowledgments: We thank Alessio Moneta, Tom Stanley, participants at the Meta-Analysis in Economic Research Network Colloquium 2013 in Greenwich, and participants at the Empirical Workshop on Energy 2014 in Kassel for helpful comments.

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, 2006). It is always possible for researchers to select statistically significant results from those presented by sampling variability (Rosenthal, 1979). In the case of Granger causality testing, spuriously statistically significant results can be generated if the underlying Vector Autoregression (VAR) model is overfitted, which is increasingly likely the smaller the sample size. Small sample sizes are common in macroeconomic research using annual data. This overfitting bias is likely to further increase the number of false-positive results published in the literature raising the question whether most findings of Granger causality are false. In this paper, we present a meta-analysis approach for synthesizing Granger-causality test statistics that deals with both publication selection bias based on selecting significant results from sampling variability and the increased false-positive rate generated by overfitted models.

Granger causality is a widely used concept in various fields of economics such as monetary policy (Mittnik and Semmler, 2013; Lee and Yang, 2012; Assenmacher-Wesche and Gerlach, 2008; Aksoy and Piskorski, 2006), finance and economic development (Ang, 2008), and energy economics (Ozturk, 2010), and it has also been increasingly applied in other scientific disciplines, such as climate change (Stern and Kaufmann, 2014), and neuroscience (Bressler and Seth, 2011). However, Granger causality test statistics are very sensitive to the chosen lag length for the underlying VAR model (e.g. Zapata and Rambaldi, 1997). Given that the true lag length is unknown, uncertainty about which lag length to select leaves researchers with substantial uncertainty about the validity of Granger causality tests. 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, 1973) and the Bayesian information criterion (BIC) (Schwartz, 1978). However, these information criteria have a known tendency to overestimate and underestimate, respectively, the true lag length (Ozcicek and McMillin, 2010; Hacker and Hatemi-J, 2008; Nickelsburg, 1985; Lütkepohl, 1985). Overfitted and underfitted models also tend to lead to overrejection and underrejection of the null hypothesis of Granger non-causality compared to the rejection rate of a model estimated with the true lag length (Zapata and Rambaldi, 1997). Overfitting and underfitting are particularly prevalent in small samples

(Gonzalo and Pitarikis, 2002), which are common in macroeconomic research using annual data. In comparison to microeconomics, reliable causal inference in macroeconomics is in any case hindered by the difficulty of applying experimental and quasi-experimental designs in macroeconomics (Stock, 2010) and finding valid instrumental variables is often difficult (Bazzi and Clemens, 2013). Hence, addressing the potential biases introduced by lag length selection in Granger causality testing is important in strengthening causal interpretations in macroeconomics.

Many approaches have been developed to improve the likelihood 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).

Considering the incentive system in academic publishing that rewards statistically significant results and compliance with the preconceptions of reviewers (Frey, 2003; Glaeser, 2006; Brodeur *et al*., 2013) means that dealing with overfitting bias is even more important. Overfitted VAR models and the corresponding overrejection of the non-causality hypothesis can be used to consciously or unconsciously obtain statistically significant Granger causality tests even if genuine Granger causality is absent. Overfitting bias is specific to the analysis of time series and comes in addition to more general biases such as the selection of statistically significant results from sampling variability by estimating the effect of interest, for example, for varying countries, time spans or data sources.[1]

We address these biases using meta-analysis. Meta-regression models can synthesize the Granger causality tests from many primary studies in order to 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). In this approach, the regression coefficients of primary studies are regressed on primary study characteristics, such as the country under investigation or the

[1] Publication bias is often defined as the selection of statistically significant results from sampling variability. This definition is based on experimental designs where sampling variability may be the main source of variation in results (Rosenthal, 1979). Publication (selection) bias in observational research, however, can exploit more sources of variation, such as omitted-variable biases. Therefore, we use publication selection bias more generally to define the selection of specific results for publication, e.g. statistically significant or theory-confirming results, not necessarily based on sampling variability alone.

estimation technique employed. Subsequently, meta-regression analysis was further developed to identify genuine empirical effects while controlling for publication selection bias based on the selection of statistically significant and theory-confirming results from sampling variability (Stanley, 2008). 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 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 test 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 overfitting bias discussed above need to be taken into account in developing a meta-regression approach that is suitable for Granger causality test statistics.

Our proposed meta-regression model for Granger causality test statistics regresses the probit-transformed p-value of the original test statistics on a constant, the square root of the degrees of freedom in the primary regressions, and control variables that include the selected lag length from the primary studies. Using Monte Carlo simulations, we show that the overfitting bias occurs in many scenarios that are likely to be prevalent in macroeconomics. We also simulate empirical literatures that are additionally distorted due to primary authors searching for statistically significant and theory-confirming results. We show that the suggested meta-regression models can identify the presence and absence of genuine Granger causality even if genuine Granger causality is absent but the primary literature only includes statistically significant Granger causality test statistics due to these biases. The suggested model that controls for overfitting bias systematically outperforms a basic model that does not control for overfitting bias.

The development of this meta-regression model is motivated by the large and inconclusive literature that tests Granger causality between energy use and economic output. We show that the presence of excess significance in this literature can be largely explained by overfitting bias rather than the presence of genuine Granger causality. These findings highlight how as a result of overfitting bias a literature can appear to provide evidence for Granger causality

when actually Granger causality appears to be absent. Understanding the mechanisms at work is important for formulating evidence-based policy and our results show that false-positive findings can be easily derived implying misleading policy implications.

Section 2 of the paper discusses testing for Granger causality, overfitting bias, and the meta-regression models. Section 3 describes the design of our simulations and Section 4 presents the results. Section 5 provides a discussion of the results and Section 6 applies the meta-regression models to the literature on energy use and economic output. Section 7 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, 42).

Testing for Granger causality requires knowledge about the properties of the time series under consideration. If the time series are integrated, Wald test statistics for Granger causality in a VAR in levels follow non-standard asymptotic distributions and depend on nuisance parameters (Sims *et al*., 1990; Toda and Phillips, 1993). Whether or not time series are integrated can be tested using a variety of unit root tests. Yet, all of these tests suffer from low power in small samples. If the time series in question are indeed first order integrated (I(1)) but not cointegrated, a VAR in first differences provides a valid framework for testing Granger causality using Wald tests. Instead, if the time series are I(1) and cointegrated, a vector error correction model (VECM) is the appropriate framework for Granger causality testing. If the time series are not integrated (I(0)), Granger causality can be directly tested using a VAR in levels as the unrestricted model. Hence, there are a variety of models that can be applied to test for Granger causality and the validity of each model depends on the properties of the specific time series involved. However, pre-testing biases are introduced by

testing for the order of integration and cointegration to decide which framework should be used for Granger causality testing.

As a remedy, Toda and Yamamoto (1995) proposed a Granger causality testing procedure that avoids any pre-testing. They show that if a VAR in levels is augmented by a 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, Granger causality can be tested 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: $\Pi_1^{ab} = \Pi_2^{ab}$ ... $= \Pi_p^{ab} = 0$, 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:

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

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

2.2. Overfitting Bias

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 are most commonly used. Ignoring a constant that reflects the number of deterministic terms, the general information criterion for selecting the lag length in VAR models is:

$$IC(p) = ln\left|\hat{\Sigma}_p\right| + \frac{c_T}{T}\, p\, q^2 \quad (3)$$

where *IC(p)* is the value of the information criterion for a lag length p, $\hat{\Sigma}_p$ is the estimated covariance matrix of the $VAR(p)$ model, T is the number of observations and q is the dimension of the VAR. The true lag length p^* is estimated using:

$$\hat{p}^* = \arg\min_{\mathrm{p}} IC(p) \tag{4}$$

with $0 \leq p \leq p_{max}$. Given that $\left|\hat{\Sigma}_p\right|$ will always decrease as a result of adding further lags to the model, a penalty term is introduced, which is the second term on the right hand side of (3). Hence the IC only recommends adding an additional lag if the decrease in the log likelihood exceeds the penalty term applied. The AIC sets the deterministic penalty term to $c_T = 2$, whereas the BIC uses $c_T = \ln T$.

Nielsen (2006) shows that $\hat{p}^*$ is a consistent estimate of the true lag length if the BIC is used, whereas the AIC has a positive limiting probability of overfitting. 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 of a few decades of quarterly or annual data. Hence, there is usually a small to moderate number of observations. In addition to this, it is questionable whether very long time series can be used in a single econometric model due to potential changes in the underlying data generating process (DGP), for example due to changes in economic policy (Lucas, 1976).

Accordingly, it is the performance of IC's in small and moderate sample sizes that matters in applied macro-econometrics. At these sample sizes, the frequency with which the AIC and the BIC select the true lag length p^* depends heavily on the specific DGP. For instance, Hacker and Hatemi-J (2008) illustrate how the size of the coefficient of the last lag in the VAR model influences the ability of the IC to select p^*. Furthermore, Nickelsburg (1985) shows that the frequency of over- or underfitting also depends on whether the lag coefficients decline exponentially, oscillate, sharply increase, or exhibit dampened oscillations. Although, the exact frequency distribution may vary with respect to the specific DGP, systematic patterns can be identified when IC's are used. Based on (3), the probability to overfit a $VAR(p^*)$ model by h lags is:

$$P[IC(p^*) > IC(p^* + h)] = P\left[ln\left|\hat{\Sigma}_{p^*}\right| - ln\left|\hat{\Sigma}_{p^*+h}\right| > \frac{c_T\, p^*\, q^2\, h}{T}\right]. \tag{5}$$

If there are few degrees of freedom (*df*), the sampling variability of $\hat{\Sigma}_p$ will be large. As a result, the variance of $ln|\hat{\Sigma}_{p^*}| - ln|\hat{\Sigma}_{p^*+h}|$ can become large while the penalty term is not affected by the 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:

$$P[IC(p^*) > IC(p^* - h)] = P\left[ln|\hat{\Sigma}_{p^*-h}| - ln|\hat{\Sigma}_{p^*}| < \frac{c_T\, p^*\, q^2\, h}{T}\right]. \quad (6)$$

The potentially large variance of $ln|\hat{\Sigma}_{p^*-h}| - ln|\hat{\Sigma}_{p^*}|$ due to sampling variability for low *df* 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.

Overall, the probability of overfitting and underfitting the VAR model increases with decreasing *df* with overfitting more likely to occur using the AIC and underfitting when using the BIC. These patterns have been shown in simulations of a variety of DGPs including stable and unstable VARs under situations of homoscedasticity and ARCH (Hacker and Hatemi-J, 2008), VARs with high lag lengths (Nickelsburg, 1985) and low lag lengths (Lütkepohl 1985), as well as symmetric and asymmetric lag lengths (Ozcicek and McMillin 2010).

Overfitted VAR models tend to overreject the null hypotheses of Granger non-causality and, analogously, underfitted VAR models tend to underreject 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 in the absence of genuine Granger causality. We denote this bias in Granger causality testing as overfitting bias. Given the pressure to publish statistically significant results, results that are statistically significant due to overfitting bias will be more likely to be published than insignificant results and researchers may choose the information criterion that generates such seemingly significant results. Underfitted VAR models lead to less significant Granger causality tests - we denote this bias underfitting bias. Though it will result in type II errors, this is of less importance as researchers will be less likely to publish these results.[2] Given overfitting and the incentive

[2] Of course, if theory predicts a lack of causality then researchers may be happy to publish these results.

system of academic publishing, researchers are left with uncertainty about the reliability of inferences obtained by Granger causality tests.

2.3. Meta-Granger Causality Model

In the absence of genuine Granger causality, even if the VAR model is estimated with the correct lag length, false-positive findings of Granger causality will still appear by chance. If authors select for statistically significant results from the results offered by sampling variability, the empirical literature will be distorted. Researchers can select results from estimates for a variety of countries, different time spans, or data sources. In the worst case, the published literature may consist of just the 5% of results where the null hypothesis was rejected, whereas the remaining 95% of results remain unpublished (Rosenthal, 1979). The following basic meta-regression model for Granger causality test statistics (Bruns *et al*., 2014) tests for the presence of genuine Granger causality in the presence of publication selection bias based on sampling variability but not on overfitting or underfitting biases:

$$z_i^{gc} = \alpha_B^{gc} + \beta_B^{gc}\sqrt{df_i} + \varepsilon_i^{gc} \qquad (7)$$

where df_i is the number of degrees of freedom of a single equation of the VAR used in primary study i and $z_i^{gc} = \Phi^{-1}\left(1 - \pi_i^{gc}\right)$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution and π_i^{gc} is the p-value of study i. The direction of Granger causality is given by $g = 1, \ldots, (1-q)$ denoting the independent variable in equation $c = 1, \ldots, q$ of the VAR so that, for example, $g = 1$ and $c = 2$ represents Granger causality from the first independent variable to the dependent variable in the second equation of the VAR.[3] Larger values of z_i^{gc} indicate smaller p-values and, consequently, higher levels of statistical significance. If there is no genuine effect the probit transformation of the p-values results in a normal distributed dependent variable with mean zero and hence ε_i^{gc} has desirable properties for a regression residual. The distribution of z^{gc} is more complicated in the presence of genuine Granger causality and will be discussed below. In the presence of genuine Granger causality, the level of statistical significance should increase as df_i increases ($\beta_B^{gc} > 0$). Conversely, in the absence of genuine Granger causality, df_i should be

[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 6 of this article or Bruns *et al*. (2014) for more details. In following we assume for simplicity that each primary study reports Granger causality test statistics from a single estimated VAR model. In practice studies often report the results from multiple samples and specifications. We take this into account in the empirical example.

unrelated to the levels of statistical significance. In the presence of publication selection bias based on sampling variability, 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 even if the primary literature exclusively consists of statistically significant results generated from sampling variability. Simulations show that meta-regression models of this type can control for publication selection bias that is based on sampling variability (Stanley, 2008; Bruns, 2013) and, thus, $H_0: \beta_B^{gc} \leq 0$ tests for the presence of genuine Granger causality in the presence of this publication selection bias. We refer to this model as the basic meta-regression model for Granger causality tests.

As discussed in Section 2.2, overfitting bias might be used to consciously or unconsciously search for statistically significant Granger causality tests. Meta-regression analysis can help to identify the presence of genuine Granger causality in the presence of overfitting bias. Overfitting bias leads to large values of z^{gc} compared to the values of z^{gc} that we can expect for the true lag length and these large values of z^{gc} are more common for small values of df. 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 in the presence of overfitting bias. We suggest controlling for the underlying lag length of the VAR model in the meta-regression model to account for this source of bias:

$$z_i^{gc} = \alpha_E^{gc} + \beta_E^{gc} \sqrt{df_i} + \gamma^{gc} p_i + v_i^{gc} . \quad (8)$$

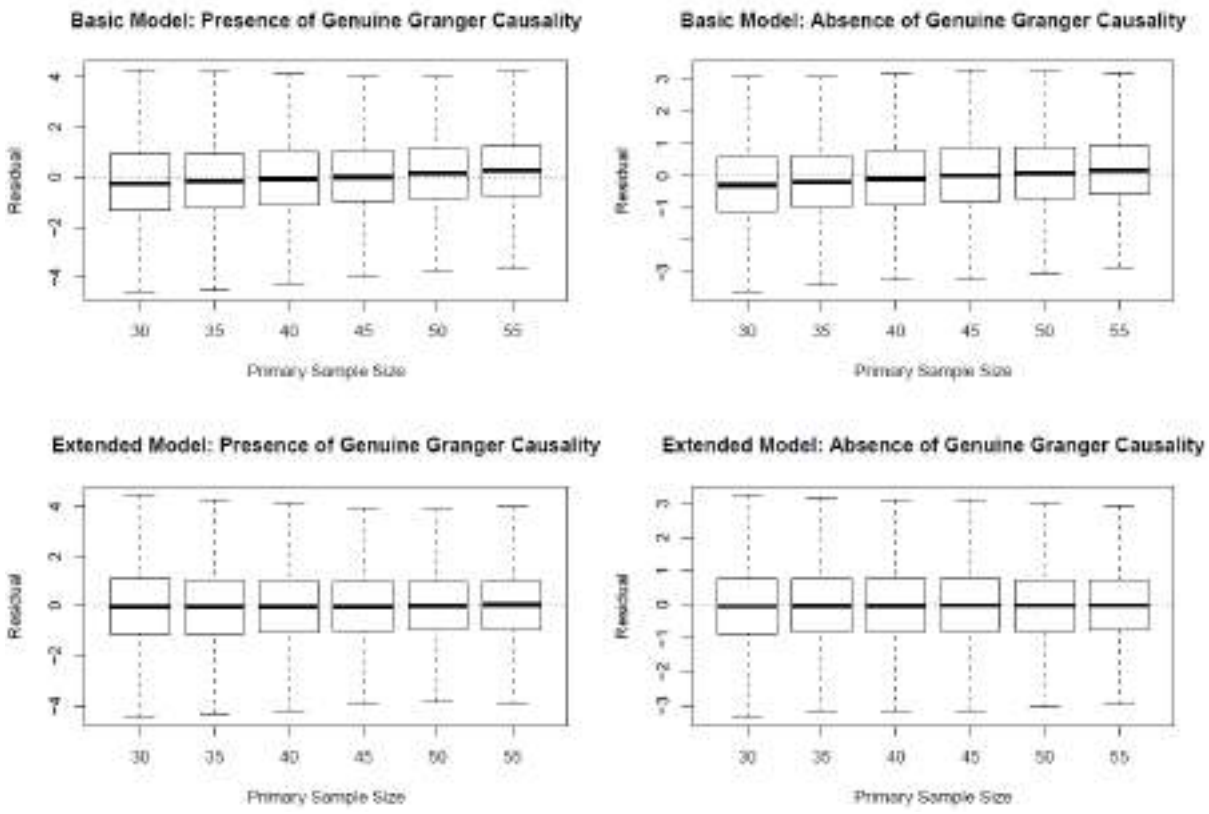
We refer to this model as the extended meta-regression model for Granger causality tests. In the presence of publication selection bias that is based on sampling variability and overfitting bias, genuine Granger causality is tested by $H_0: \beta_E^{gc} \leq 0$.

We illustrate the distribution of z^{gc} in the presence and absence of genuine Granger causality by using a small Monte Carlo simulation. We utilize the first DGP from Section 3, which will be discussed later in more detail. This DGP is a bivariate VAR process with Granger causality in one direction only and a lag length of three. In this bivariate VAR, the presence of Granger causality is mirrored by $gc = 11$ whereas the absence of Granger causality is mirrored by $gc = 12$. We generate data using this DGP for 11 different primary sample sizes $(30, 32, 35, 37 \ \ldots, 55)$ that are typical in the macroeconomic analysis of annual time series.

For each primary sample size 10,000 pairs of time series are generated following the procedure outlined in Section 3. We estimate for each of the primary sample sizes 10,000 VAR models choosing the lag length using the AIC with a maximum lag length of five. We use the Toda-Yamamoto procedure to test for Granger causality resulting in 110,000 Granger causality tests in each potential direction of causality. Finally, we fit the basic and extended meta-regression models to all 110,000 Granger causality tests – one meta-regression for each direction of causation.

Figure 1 shows the residual distributions for both the basic and extended meta-regression models grouped by primary sample size. This figure indicates evidence for non-linearity in the basic meta-regression model, which is resolved by the use of the extended meta-regression model. For the extended meta-regression model, the residuals are symmetrically distributed around zero both in the absence of and presence of genuine Granger causality.

Figure 1: Boxplots of Residuals by Primary Sample Size

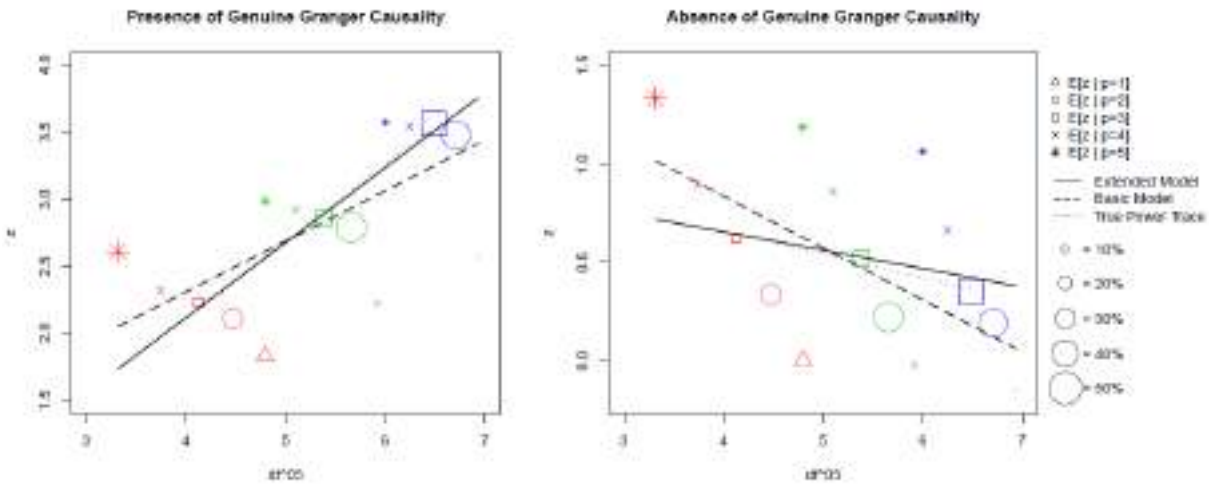


Notes: The box represents the interquartile range and the whiskers extend to the largest data point within 1.5 times the interquartile range.

The main source of non-linearity in the basic meta-regression model is overfitting bias, which generates large values of z^{gc} for small values of df. As a result, the estimate of β_B^{gc} is biased downwards obscuring the true relation between z^{gc} and $\sqrt{df}$. Figure 2 illustrates how β_B^{gc} is biased downwards due to overfitting bias and how the extended meta-regression model corrects for this bias. This small simulation also reveals that the relation between $\sqrt{df}$ and $z^{gc=12}$, that is the direction where Granger causality is absent, is negative for the true model with three lags where we would actually expect $E[z^{12} \mid df] = 0$. This negative relation indicates that the Toda Yamamoto test generates false-positive findings of Granger causality if the sample size is small. As a result, we can expect $\beta_E^{gc} < 0$ in the absence of genuine Granger causality for samples sizes that are typical of the macroeconomic analysis of annual time series.

Given that β_B^{gc} is biased downwards and genuine Granger causality is characterized by $\beta_B^{gc} > 0$, we can expect that overfitting bias means that the basic model has lower power than the extended model. In addition, the size of the basic model can be expected to be lower than the nominal significance level of 5% and lower than that of the extended model.

Figure 2: Impact of Overfitting Bias on the Meta-Regression Models



Notes: The average z^{gc} values per chosen lag length ($E[z^{gc}|p]$ with $p = 1,2,3,4,5$) are represented as function of $\sqrt{df}$ for three different primary sample sizes: $\mu = 30$ (red), $\mu = 40$ (green), and $\mu = 55$ (blue). The size of the symbols mirrors for each of the three primary sample sizes the percentage of the primary studies that choose the respective lag length. The relation between z^{gc} and $\sqrt{df}$ for the true lag length of three ($p = 3$) is represented by the dotted line (true power trace). The slope of the Basic Model (β_B^{gc}) is biased downwards compared to the true relation between z^{gc} and $\sqrt{df}$. The extended model is plotted for a lag length of three.

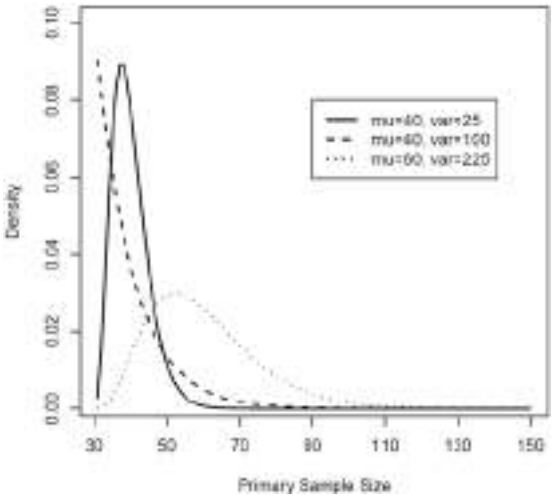
3. Simulation Designs

3.1. No Publication Selection

In this subsection, we describe a simulation that is intended to show how well the basic and extended meta-regression models perform in the presence of overfitting bias in Granger Causality testing if the authors of primary studies do not select for specific results.

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 n_i for each primary study 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 next 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 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 and 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 Figure 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 Figure 3 and provides a 10% (90%) quantile of 34 (47). Quarterly time series provide more observations but are usually available for fewer years. Quarterly time series often start around 1990 and if the meta-analyst considers again studies of the last 15 years the primary sample sizes range between 40 and 80. Figure 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).

Figure 3: Distributions of Primary Sample Sizes



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 and the roots of the companion matrix of all four DGPs lie on or outside the unit circle so that they are all non-stationary. 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.

DGP1a is a non-cointegrated bivariate VAR process. 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 coefficient. DGP1b has a larger casual effect than DGP1a but is otherwise identical. This variation in the value of the causal effect allows us to evaluate the performance of meta-regression models for different sizes of causal effects and consequent signal to noise ratios. DGP2a is a cointegrated bivariate VAR process. DGP2a deviates from DGP1a only in the second coefficient matrix, which results in a reduced rank long-run coefficient matrix as is necessary to achieve cointegration. DGP2b has a larger causal effect than DGP2a but is otherwise the same. 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 eight DGPs in total.

For each primary study i, we draw three starting values for X and Y from standard normal distributions and generate $n_i + 50$ observations according to the respective DGP. Afterwards we delete the first 50 observations to avoid any dependence on the starting values.

Each primary study applies an information criteria (AIC or BIC) to a VAR in levels to determine the optimal lag length ($p = 1, \dots, 5$). Subsequently, the lag length is augmented with the maximum order of integration of one leading to a theoretical minimum degrees of freedom of 11. Finally, each primary study applies a Wald test to the lags of the independent variable ignoring the augmented 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 ($\#s * \#\mu * \#\sigma^2 * \#DGP * \#IC$).

Table 1: Overview of Data Generating Processes

Name	VAR 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	$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 1.5 & 0.8 \\ 0 & 1.5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.25 & -0.4 \\ 0 & -0.25 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$

3.2. Theory-Confirmation Bias

We also examine the case where researchers search for theory-confirming and statistically significant results. 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, authors may search for results that confirm these theoretical presumptions.

We generate the primary sample sizes, n_i, as described in section 3.1. 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 the more statistically significant. Moreover, we consider that researchers conducting $h\%$ (where $h = 0,25,50,75,100$) 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 if they do not find a result that is significant at the conventional level of 5% they also search further samples of data (from other countries or time periods) until they find Granger causality from Y to X that is statistically significant at the 5% 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 5% level. This gives further opportunities to generate apparently significant results due to selection from sampling variability and overfitting bias.

As a result, the primary literature is composed of $h\%$ primary studies with statistically significant Granger causality tests from Y to X due to publication selection bias based on sampling variability and overfitting bias. 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 ($\#s * \#\mu * \#\sigma^2 * \#DGP * \#h$).

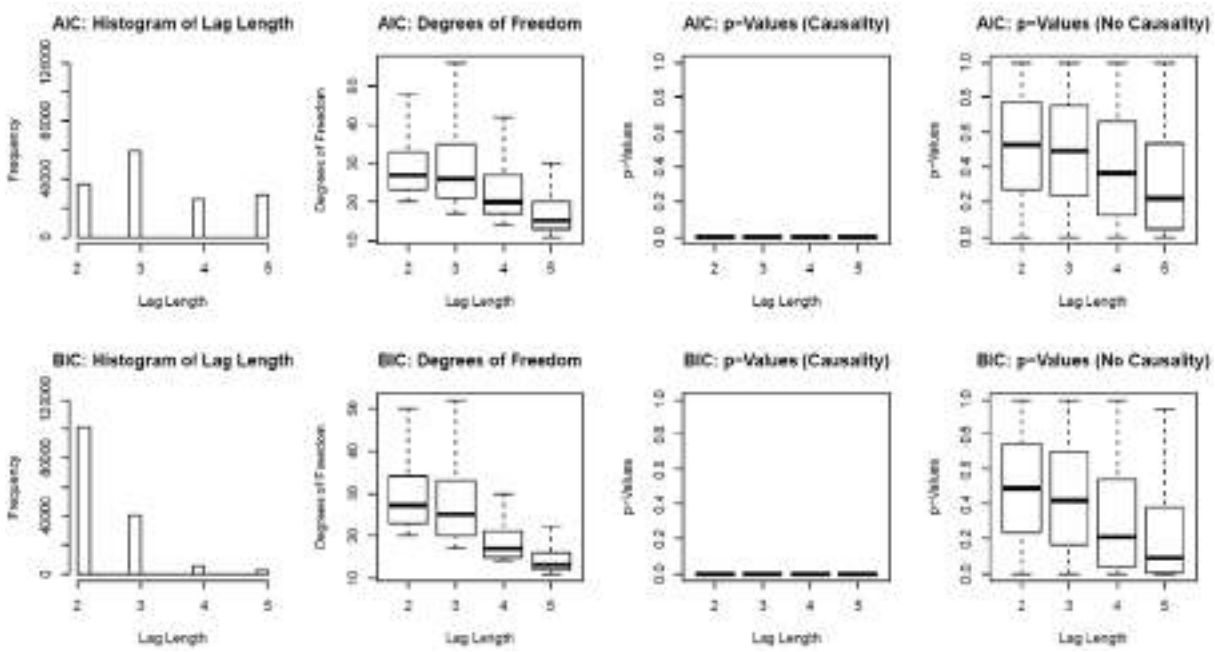
4. Results

4.1. No Publication Selection

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 and they tend to choose the correct lag length for larger degrees of freedom. 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 5%. 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 indicating overfitting bias. Underfitted VAR models have p-value distributions with a larger mean compared to the VAR model with the true lag length indicating underfitting bias. Figure 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 if the AIC is used overfitting bias occurs frequently in a variety of scenarios that mirror actual research in empirical macroeconomics.

Figure 4: Prevalence of Overfitting Bias for DGP2a

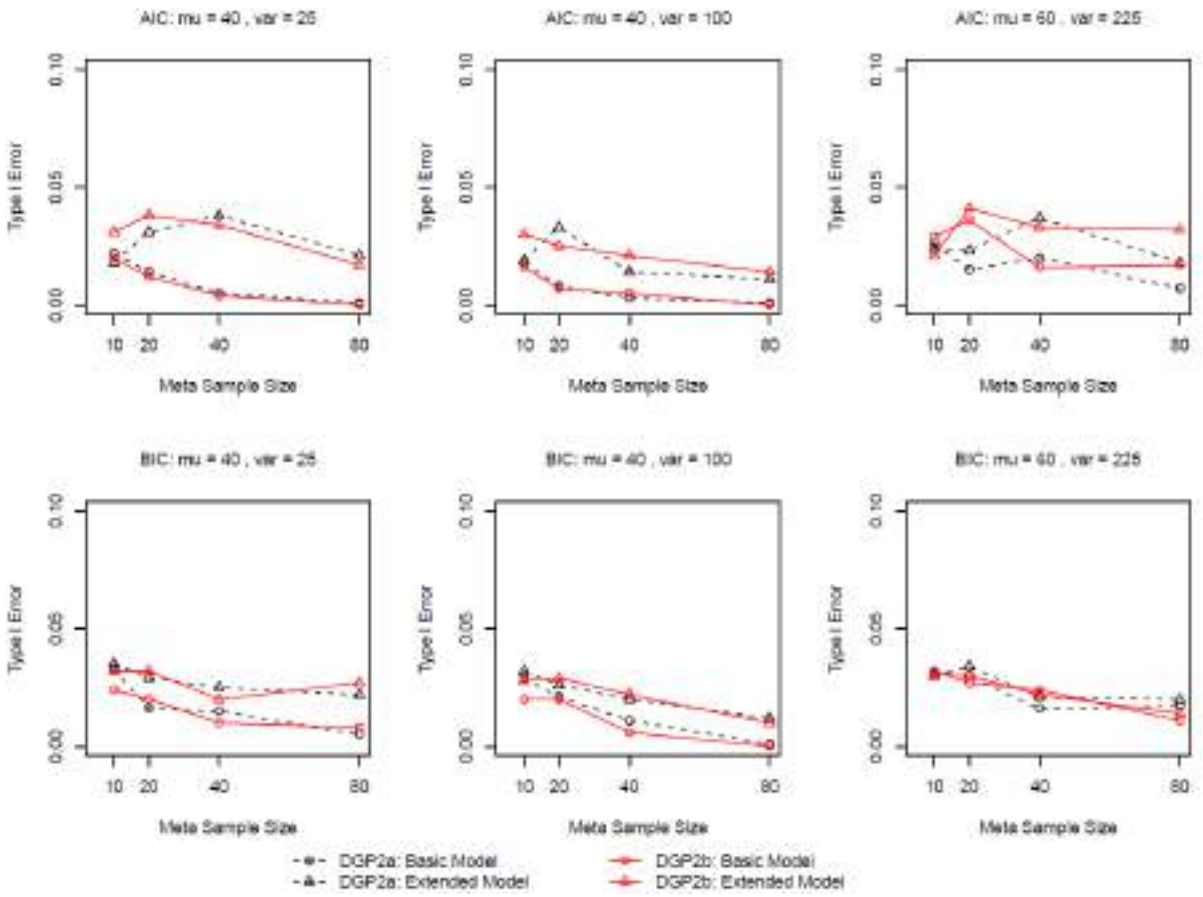


Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) resulting in 150,000 observations using a primary sample size distribution with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = \mathrm{I}$. 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.

Figure 5 shows how the type I errors of both meta-regression models vary with the meta-analysis sample size for DGP2 (the cointegrated DGP). The type I errors of the basic meta-regression model are mostly smaller than the size 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 5%. This shows that β_E^{gc} is still biased downwards because the underlying distribution of the Toda-Yamamoto test statistics depends on the degrees of freedom as shown in Figure 2. DGP1 shows the same patterns as DGP2 (See Appendix A1 for DGP1).

Figure 5: Type I Errors of Meta-Regression Models for DGP2a and DGP2b

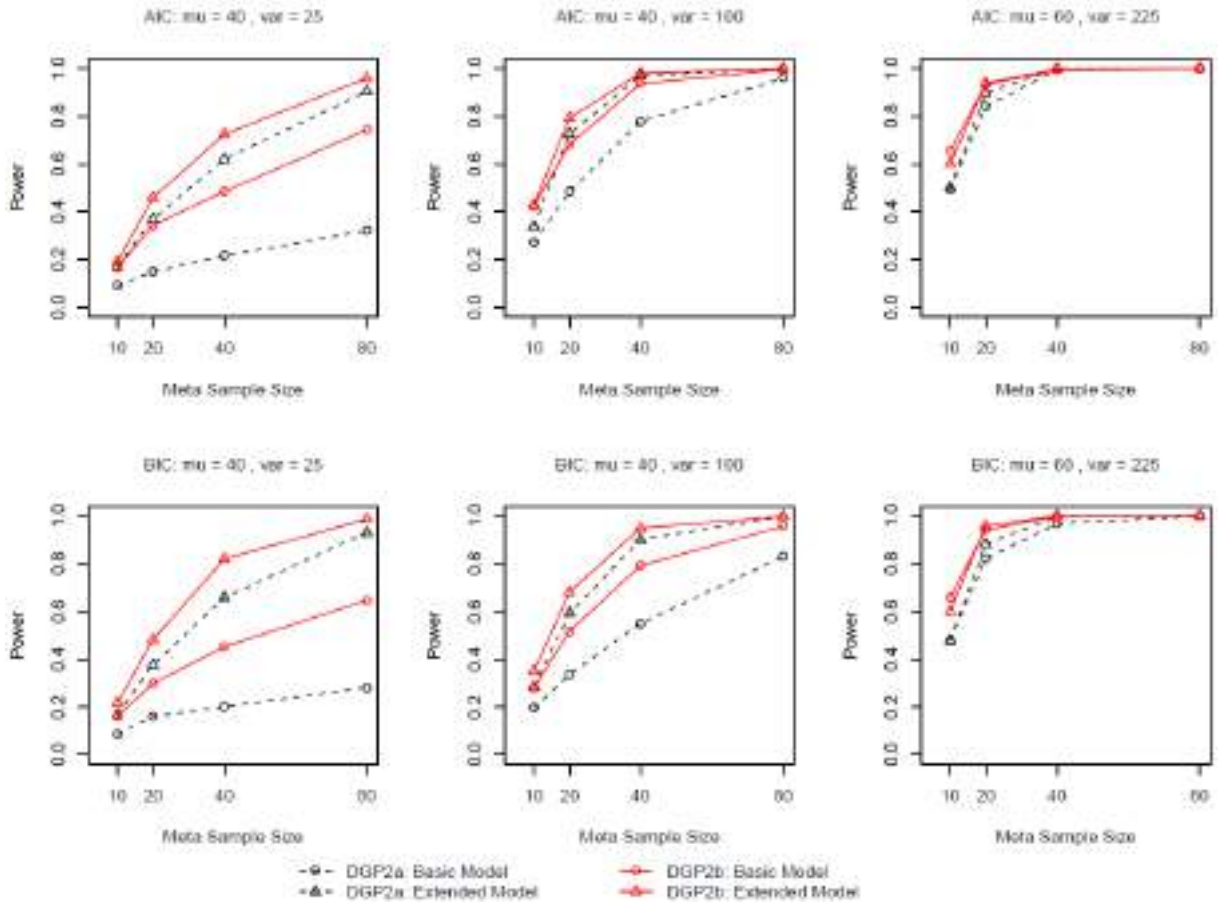


Notes: Type I errors of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = \mathrm{I}$ are reported if the AIC (upper row) or the BIC (lower row) is used for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Figure 6 shows the power of both meta-regression models in identifying genuine Granger causality in relation to the meta-analysis sample size for DGP2. For very small meta-analysis 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 $H_0: \beta_B^{gc} \leq 0$ 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 overfitting bias is actually more prevalent for the AIC. The reason is that the use of BIC leads to overfitted VAR models with exceptionally small df. The difference between the two meta-regression models decreases if the VAR errors are correlated.

Figure 6: Power of Meta-Regression Models for DGP2a and DGP2b



Notes: Power curves of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (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).

4.2. Theory Confirmation Bias

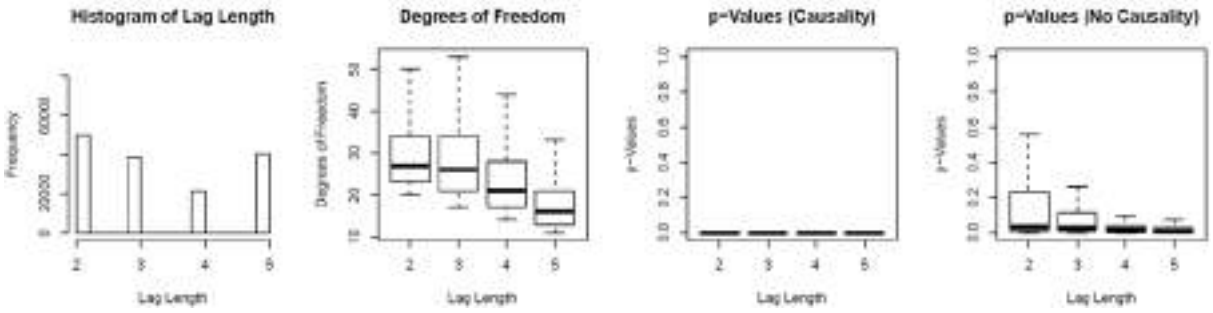
In the second case, the primary study authors search for statistically significant and theory-confirming results, that is Granger causality from Y to X, where genuine Granger causality is actually absent. Figure 7 shows that overfitted VAR models are more prevalent in this case, indicating that overfitting bias was used in addition to sampling variability to obtain statistically significant Granger causality tests for Y causes X. A large amount of excess significance is present for Y causes X, indicating how distorted an empirical literature could become.[4]

The type I errors of both meta-regression models are again well below the nominal significance level of 5%. Figure 8 shows how they vary with the degree of publication selection for DGP2. 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 publication selection, the type I errors of the basic model are even smaller indicating the increased presence of overfitting bias that increases 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 publication selection except when of 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 publication selection for DGP2. Publication selection based on sampling variability and overfitting bias 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 publication selection bias. DGP1 shows the same patterns as DGP2 (see Appendix A2).

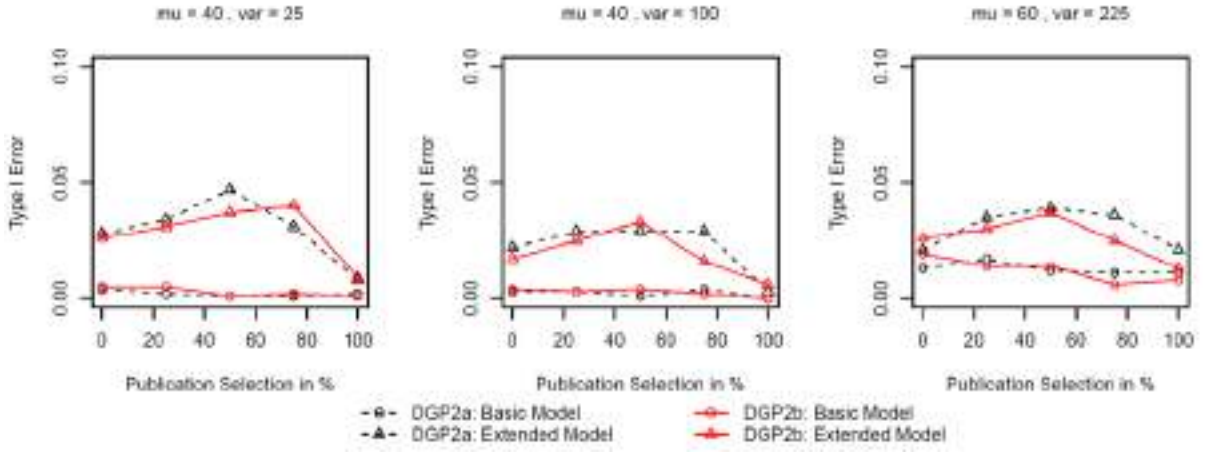
[4] We also analyzed a case in which primary studies select for any statistically significant Granger causality test irrespective of the direction of causality. In this case almost no publication selection bias occurs as genuine Granger causality is present in all DGPs and this genuine Granger causality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

Figure 7: Prevalence of Overfitting Bias for DGP2a in the Presence of Theory Confirmation Bias ($h = 75$)



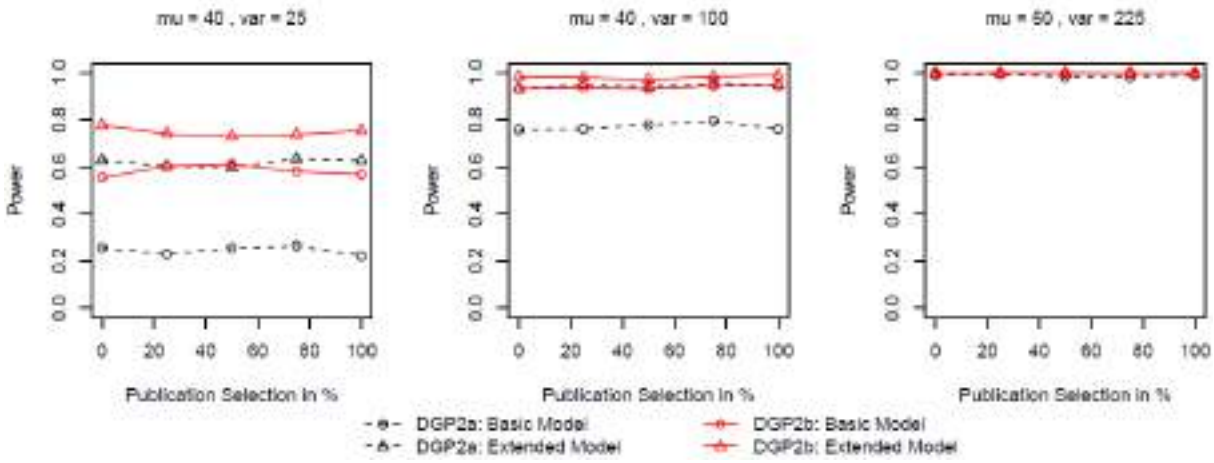
Notes: See caption of Figure 4 for further details.

Figure 8: Type I Errors of Meta-Regression Models for DGP2a and DGP2b in the Presence of Theory-Confirmation Bias



Notes: Type I errors of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = \mathrm{I}$ are reported as function of publication selection ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Figure 9: Power of Meta-Regression Models for DGP2 in the Presence of Theory-Confirmation Bias



Notes: Power curves of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (triangles) for DGP2a (black) and DGP2b (red) with $\Omega = \mathrm{I}$ are reported as function of publication selection ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

5. Discussion

We show that overfitting bias occurs in the small and moderate sample sizes that are common in macroeconomic research. This overfitting bias can lead to excess significance in an empirical literature irrespective of the presence or absence of genuine Granger causality and it hampers inference on the presence of genuine Granger causality using meta-regression models. We show that the extended meta-regression model can control for this overfitting bias and adequately distinguishes between the presence and absence of genuine Granger causality, even if all Granger causality tests are statistically significant in the primary literature due to biases.

It is well known that the AIC and the BIC tend to overfit and underfit VAR models in small samples (e.g. Gonzalo and Pitarikis, 2002). It has also been shown that overfitted and underfitted VAR models tend to overreject and underreject the null hypotheses of Granger non-causality (e.g. Zapata and Rambaldi, 1997). We contribute to these separate findings by providing simulation results that cover the complete process of Granger causality testing including lag length selection by information criteria and publication selection biases. We can show that overfitting and underfitting indeed frequently occur in simulated data that mirrors

empirical research in macroeconomics and that these misspecifications lead to underrejection and overrejection of the null hypotheses of Granger non-causality. If authors search for statistically significant Granger causality tests where genuine Granger causality is actually absent, overfitting bias can be the main source of false-positive findings.

We show that overfitting bias can be controlled for if the lag length of the underlying VAR model is introduced in the meta-regression model. The simulation results reveal that the basic meta-regression model has problems with detecting 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 bias into account. Economic effects are often small, highlighting how important the correction for overfitting bias is to reliably distinguish the presence and absence of genuine Granger causality. We also show that these findings hold for cases where primary authors select for statistically significant and theory-confirming Granger causality tests. In this case, false-positive findings of Granger causality in the primary literature can also be obtained by selecting from sampling variability. Meta-regression models have been shown to be robust with respect to this bias (Stanley, 2008; Bruns, 2013). Even if all Granger causality tests in an empirical literature are statistically significant, though genuine Granger causality is absent, both meta-regression models have a type I error rate of less than 5%.

As a result, meta-regression models can be used to improve the reliability of inferences in Granger causality testing. In microeconomics, the use of randomized experiments, instrumental variable techniques and other approaches to identifying causal effects has improved the quality of causal inference tremendously (Angrist and Pischke, 2010). However, in macroeconomics the applicability of experiments is limited (Stock, 2010) and finding valid instruments is often hard (Bazzi and Clemens, 2013). Meta-regression models allow empirical economists to gain more certainty about Granger causality between two variables despite the uncertainty introduced by overfitting bias.

6. Empirical Application: The Energy-Growth Literature

6.1. Background and Data

This paper is motivated by the very large but inconclusive empirical literature that investigates the relationship between energy consumption and economic output using Granger causality tests (Stern and Enflo, 2013). We develop improved meta-regression

models to identify the presence or absence of genuine Granger causality in this literature. Our analysis is based on the meta-analysis of Bruns *et al*. (2014) that provides first insights in the potential publication selection biases in this literature.

We select those studies that use the Toda-Yamamoto procedure to test for Granger causality from the data set of Bruns *et al*. (2014). Appendix A3 provides an overview of these 23 studies that use the Toda-Yamamoto procedure. 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.[5]

The average z^{gc} value in the sample for energy causes growth is 0.83, which corresponds to an average *p*-value of 0.20 and the average z^{gc} value for growth causes energy is 1.03, which corresponds to an average *p*-value of 0.15. Both 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. For these bivariate specifications 19.70% are statistically significant at the 5% level for energy causes growth and 27.27% for 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 5% 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 control variables (CO_2 emissions, energy prices, labor, capital, and population) but insufficient observations to group them into separate categories.

[5] 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.

Table 2: Properties of Granger Causality Estimates

Control Variables	Obs.	Energy-Growth (p-value < 0.05)	Growth-Energy (p-value < 0.05)	Percentiles of df			Percentiles of Lag Length		
				25	50	75	25	50	75
None	66	19.70%	27.27%	28	35	38	1	1	2
Capital and Labor	41	48.78%	46.34%	12	14	21	2	3	3
Other	19	10.53%	36.84%	17	21	28.5	1	1	2

6.3. Meta-Regression Models

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 these omitted-variable biases in the primary literature, meta-regression models will also detect spurious "genuine effects". By controlling for the different VAR specifications used in the primary literature we can discuss whether a positive relation between z^{gc} and $\sqrt{df}$ is due to omitted-variable bias or not.[6]

Furthermore, the addition of control variables to the primary VAR specification can deplete the degrees of freedom increasing the probability of obtaining statistically significant Granger causality tests due to overfitting bias. In general, adding variables to the VAR model increases the penalty terms of the information criteria and decreases the probability of overfitting. But if the addition of variables is used to deplete the *df* leading to very low *df*, the increased variance of $ln\left|\hat{\Sigma}_{p^*}\right| - ln\left|\hat{\Sigma}_{p^*+h}\right|$ may exceed the increase in the penalty term implying a higher probability of overfitting (Gonzalo and Pitarakis, 2002).

We generalize the extended meta-regression model (8) to take the dependence between the Granger causality test statistics and the three primary VAR specifications into account, using the following regression:

$$z_i^{gc} = \alpha_1^{gc} + \beta_1^{gc}\sqrt{df_i} + D_{KL}\left(\alpha_2^{gc} + \beta_2^{gc}\sqrt{df_i}\right) + D_{Ot}\left(\alpha_3^{gc} + \beta_3^{gc}\sqrt{df_i}\right) + \gamma^{gc}\, p_i + \varepsilon_i^{gc} \quad (9)$$

[6] 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.

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.[7] Accordingly, $H_0: \beta_1^{gc} \leq 0$ tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ if the bivariate VAR specification was used and $H_0: \beta_1^{gc} + \beta_2^{gc} \leq 0$ tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ if capital and labor are used as control variables. We use standard errors clustered by publication to account for the dependence between the error terms of the multiple Granger causality test observations provided by most individual publications.

6.4. Results and Discussion

Table 3 presents the results of the meta-regression models for energy causes growth and *vice versa*.[8] The first columns present the basic model and the corresponding estimate of β_B^{gc} is negative indicating the presence of overfitting bias though it is only statistically significant at the 10% level for energy causes growth in a two sided test. Of course, this is statistically insignificant in a one-sided test that tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$. The second columns show the extended model. Adding the lag length as a continuous control variable leads to an estimate of β_E^{gc} that is close to zero and statistically insignificant. The coefficient of the lag length variable is as expected positive and statistically significant. The third columns show the generalized extended model (9) that tests for a positive relation between z_i^{gc} and $\sqrt{df_i}$ for each of the three categories of primary VAR specifications. For both energy causes growth and *vice versa*, we find that we cannot reject the null hypotheses $H_0: \beta_1^{gc} \leq 0$ and $H_0: \beta_1^{gc} + \beta_2^{gc} \leq 0$ indicating that the excess significance is caused by the presence of biases, particularly overfitting bias, rather than by the presence of genuine Granger causality. This is seen when we compare the estimate of the constant in the first and second regressions. In the basic model the intercept is large and significant while in the extended model it is insignificantly different to zero. The extended model shows that the value of the constant is mostly driven by the models with greater lag lengths. On the other

[7] 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 an other category.

[8] We also conducted the analysis by excluding Vaona (2010) who has the largest values of df - 127 and 130, which is 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 overfitting bias on the inference of the meta-regression models as we would expect by dropping observations with large df.

hand, even models with one lag have spuriously high z^{gc}-values as is shown by the high statistical significance of γ^{gc}. These cannot be due to overfitting bias and are presumably due to publication selection from sampling variability.

Figure 10 shows how large lag lengths are especially present for low *df* and how these tests tend to have the largest z^{gc}-values. Large lag lengths occur almost exclusively for the primary VAR specification with capital and labor and the Granger causality tests for this combination also have the highest levels of statistical significance, whereas Granger causality tests for VARs with capital and labor but smaller lag lengths for this specification tend to be insignificant. 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*, the search for statistically significant results may be facilitated by adding control variables to the primary VAR specification.

This empirical application shows that there is no genuine relation between energy use and economic output in bivariate VAR specifications or in VAR specifications with capital and labor as control variables. However, both of these VAR specifications 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.

Bruns *et al*. (2014) 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 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. It takes into account that statistically significant Granger causality tests are often obtained by large lag lengths and many control variables. The approach in this paper instead focuses specifically on lag overfitting as overfitting bias 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 overfitting bias occurs 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.

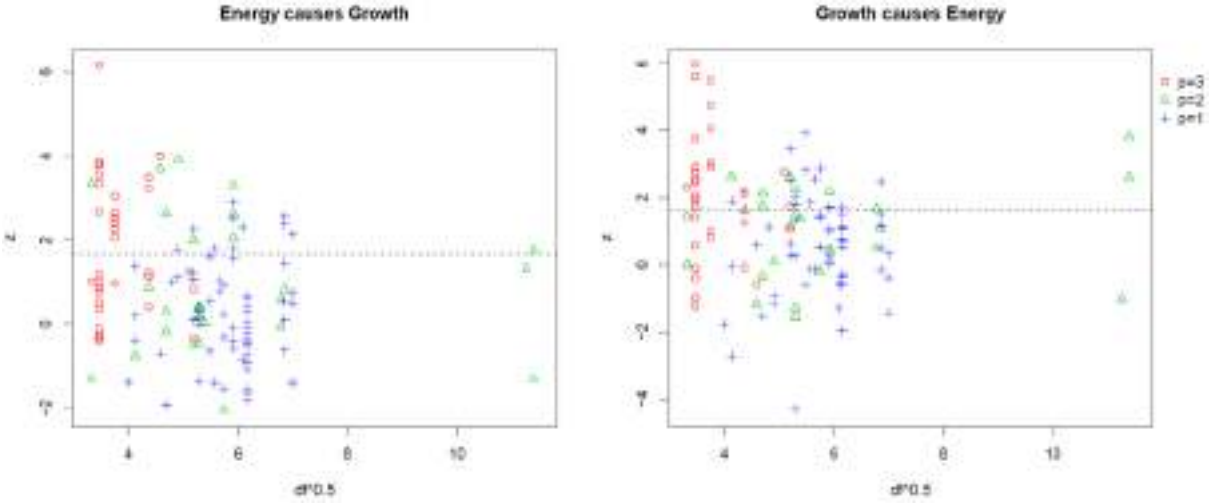
Table 3: Results of Meta-Regression Models

	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
	(0.86)	(0.86)	(3.67)	(1.11)	(1.09)	(16.51)
Df	-0.28†	-0.05	-0.06	-0.20	0.04	0.02
	(0.16)	(0.12)	(0.47)	(0.21)	(0.15)	(1.53)
Lags		0.73***	0.52		0.77**	0.70*
		(0.20)	(0.34)		(0.26)	(0.29)
KL			0.47			0.44
			(4.89)			(16.79)
KL**df*			0.04			-0.07
			(0.82)			(1.63)
Other			-1.18			-2.97
			(4.52)			(16.86)
Other**df*			0.22			0.63
			(0.70)			(1.70)
Obs.	126	126	126	126	126	126
Adj. R^2	0.06	0.17	0.18	0.02	0.13	0.12

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Notes: Bootstrapped standard errors in parentheses. We bootstrap primary studies rather than single Granger causality tests to account for the dependence of multiple Granger causality tests per primary study. Significance codes represent a two-sided t-test. One-sided t-tests representing the test for a positive relation of z^{gc} and $\sqrt{df}$ are discussed in the text.

Figure 10: Relation of Lag Length, Degrees of Freedom, and Level of Significance in the Empirical Meta Sample



Notes: 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 insignificant Granger causality tests (below).

7. Conclusions

By modeling the complete process of Granger causality testing, we show that overfitted models and the corresponding overrejection of Granger causality tests are prevalent in a variety of scenarios mirroring research in macroeconomic time series analysis. Overfitting bias leaves empirical researchers with uncertainty about the reliability of inferences. Particularly, if we consider the search for theory-confirming results, this overfitting bias is a source of excess significance even though genuine Granger causality is absent. If primary study authors adjust to the incentives of publishing statistically significant and theory-confirming results, the reliability and validity of published findings is even more uncertain and an abundance of statistically significant findings may not necessarily indicate a genuine effect.

We introduce a meta-regression model that controls for overfitting bias to help identify the source of statistically significant Granger causality tests that can be either caused by genuine Granger causality or biases. The suggested model has higher power than the basic meta-regression model and both provide adequate type I errors. These results hold for small to moderate sample sizes mirroring the analysis of annual time series in macroeconomics. The likelihood of overfitting diminishes with the larger primary sample sizes that may occur in the analysis of quarterly time series over a long time span. Searching for statistically significant results by using sampling variability or by using omitted-variables biases,

however, may be present even for larger sample sizes highlighting the need to synthesize the evidence of an entire empirical literature by means of meta-regressions to distinguish genuine effects from biases.

We apply the suggested meta-regression models 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 addition of primary control variables to the VAR models depletes degrees of freedom which increases the probability to obtain statistically significant results by overfitting bias.

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 identication. IEEE Transactions on Automatic Control, 19(6):716-723.

Aksoy, Y. and Piskorski, T. (2006). US domestic money, inflation and output. Journal of Monetary Economics, 53(2):183-197.

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.

Ang, J. B. (2008). A survey of recent developments in the literature of finance and growth. Journal of Economic Surveys, 22(3):536-576.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of Economic Perspectives, 24(2):3-30.

Assenmacher-Wesche, K. and Gerlach, S. (2008). Money growth, output gaps and inflation at low and high frequency: Spectral estimates for Switzerland. Journal of Economic Dynamics and Control, 32(2):411-435.

Bazzi, S. and Clemens, M. A. (2013). Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth. American Economic Journal: Macroeconomics 5(2):152–186.

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.

Bressler, S. L. and Seth, A. K. (2011). Wiener-Granger causality: A well established methodology. Neuroimage, 58(2):323-329.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2013). Star wars: The empirics strike back. IZA Discussion Paper No. 7268.

Bruns, S. B. (2013). Identification of genuine effects in observational research by means of meta-regressions. Jena Economic Research Papers 2013-040, Friedrich Schiller University Jena and Max Planck Institute of Economics.

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.

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

Frey, B. S. (2003). Publishing as prostitution? - Choosing between one's own ideas and academic success. Public Choice, 116(1-2):205-223.

Glaeser, E. (2006). Researcher incentives and empirical methods. NBER Technical Working Papers, 329.

Gonzalo, J. and Pitarakis, J.-Y. (2002). Lag length estimation in large dimensional systems. Journal of Time Series Analysis, 23(4):401-423.

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

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

Lee, T.-H. and Yang, W. (2012). Money-income Granger-causality in quantiles. Advances in Econometrics, 30:385-409.

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

Lucas, R. E. (1976). Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series on Public Policy, 19-46. North-Holland.

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.

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.

Mittnik, S. and Semmler, W. (2013). The real consequences of financial stress. Journal of Economic Dynamics and Control, 37(8):1479-1499.

Nielsen, B. (2006). Order determination in general vector autoregressions. IMS Lecture Notes - Monograph Series Time Series and Related Topics, 93-112. Institute of Mathematical Statistics.

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.

Ozturk, I. (2010). A literature survey on energy-growth nexus. Energy Policy, 38(1):340-349.

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.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results, Psychological Bulletin, 86(3):638-641.

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

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.

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

Sims, C. A., Stock, J. H., and Watson, M. W. (1990). Inference in linear time series models with some unit roots. Econometrica, 58(1):113-144.

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.

Stern, D. I. and Enflo, K. (2013). Causality between energy and output in the long-run. *Energy Economics*, 39:135-146.

Stern D. I. and Kaufmann, R. K. (2014). Anthropogenic and natural causes of climate change, Climatic Change, 122:257-269.

Stock, J. H. (2010). The other transformation in econometric practice: Robust tools for inference. Journal of Economic Perspectives, 24(2):83-94.

Toda, H. Y. and Phillips, P. C. (1993). Vector autoregressions and causality. Econometrica, 61(6):1367-1393.

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.

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.

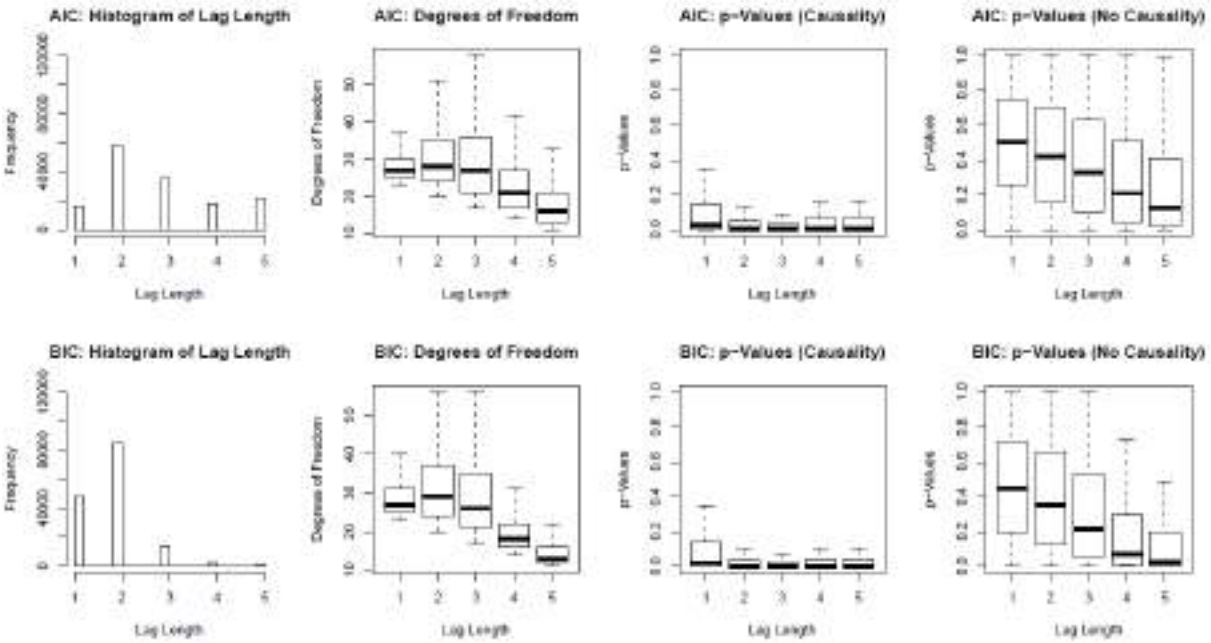
Zhang, X.-P. and Cheng, X.-M. (2009). Energy consumption, carbon emissions, and economic growth in China. Ecological Economics, 68(10):2706-2712.

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.

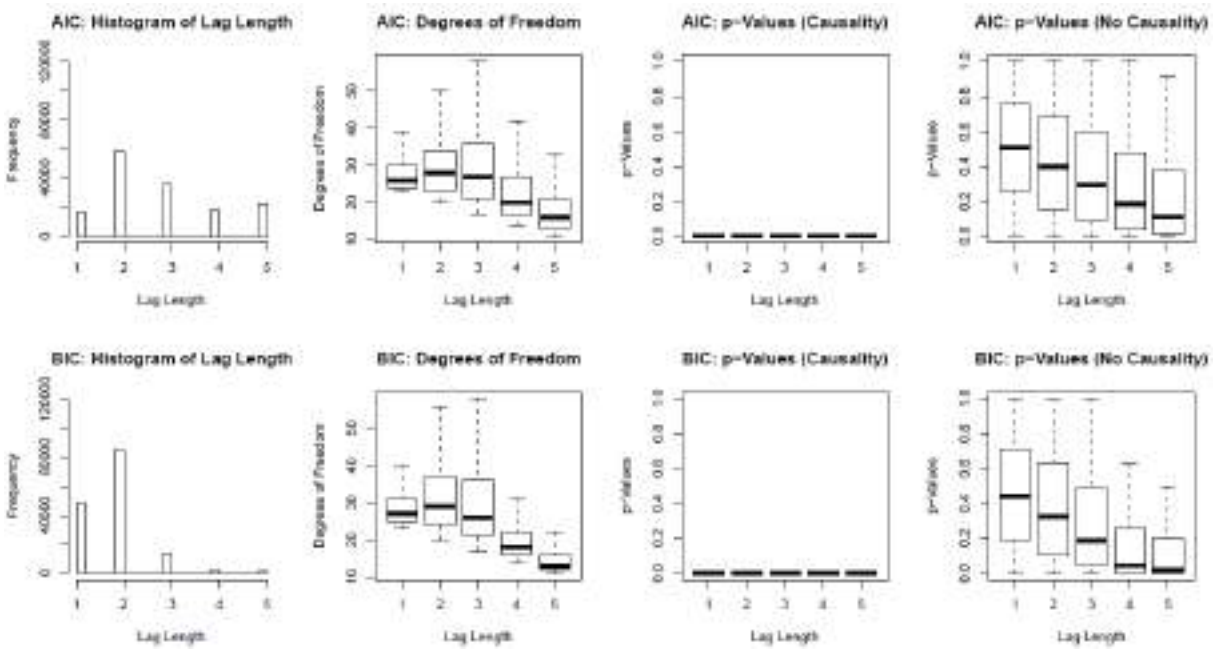
Appendix A1

Figure 4a: Prevalence of Overfitting Bias for DGP1a



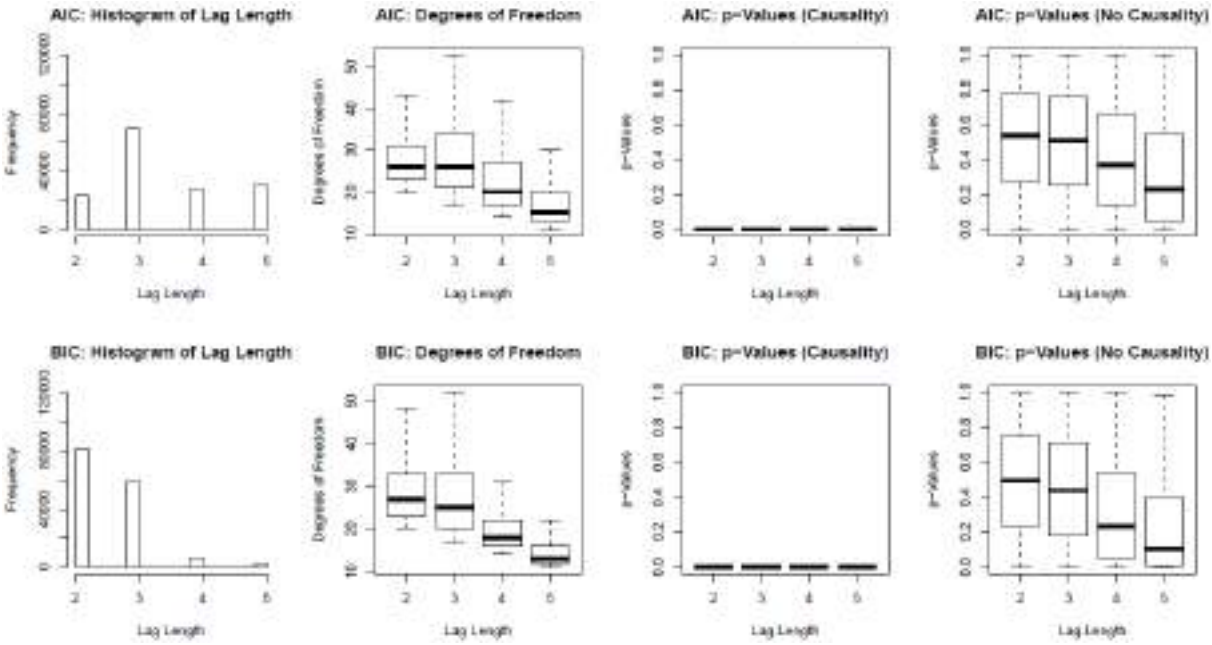
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The 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.

Figure 4b: Prevalence of Overfitting Bias for DGP1b



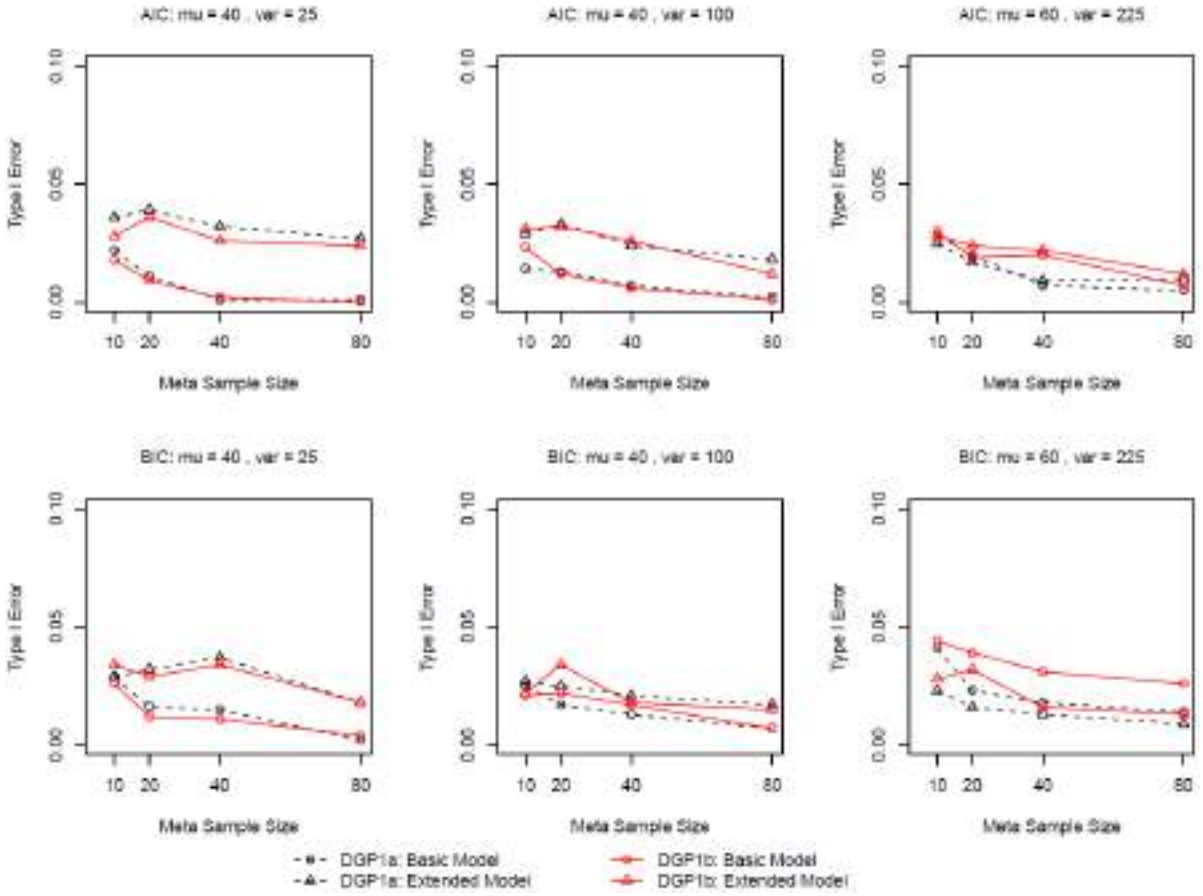
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The 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.

Figure 4c: Prevalence of Overfitting Bias for DGP2b



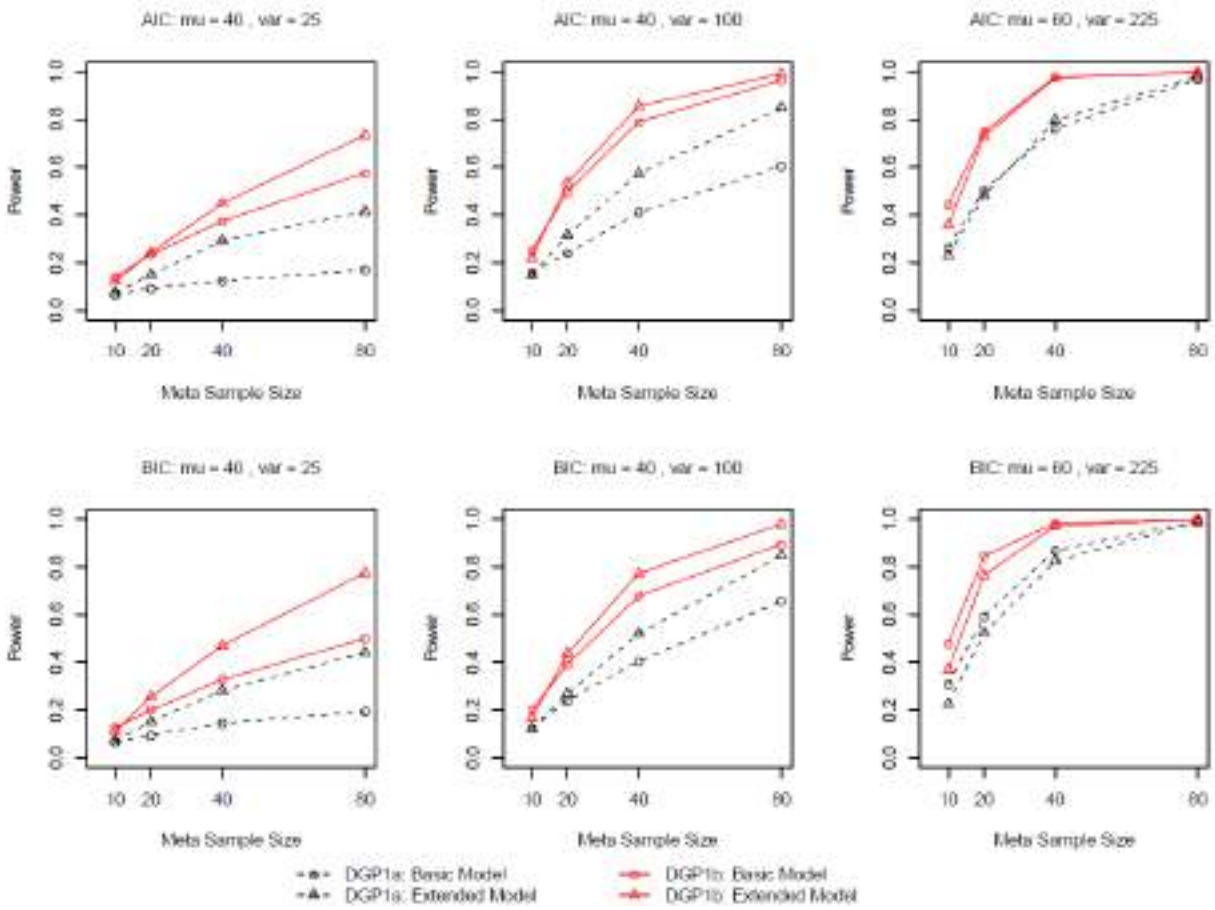
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) with $\mu = 40$, $\sigma^2 = 100$, and $\Omega = I$. The second column presents the boxplots of degrees of freedom by lag length. The 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.

Figure 5a: Type I Errors of Meta-Regression Models for DGP1a and DGP1b



Notes: Type I errors of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (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.

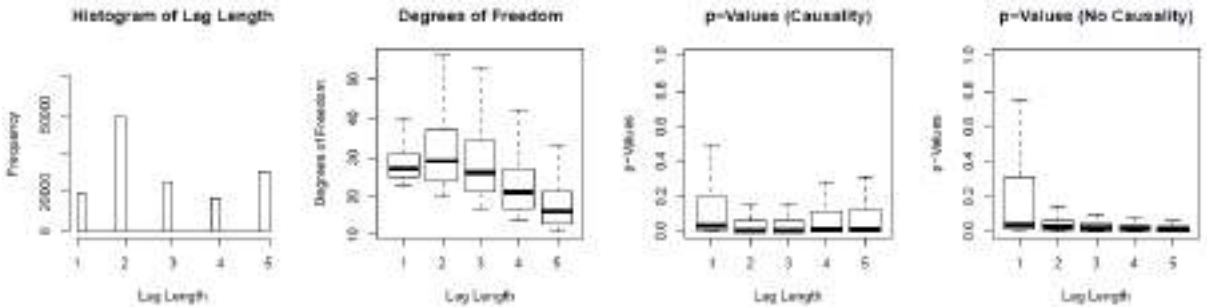
Figure 6a: Power of Meta-Regression Models for DGP1a and DGP1b



Notes: Power curves of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (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.

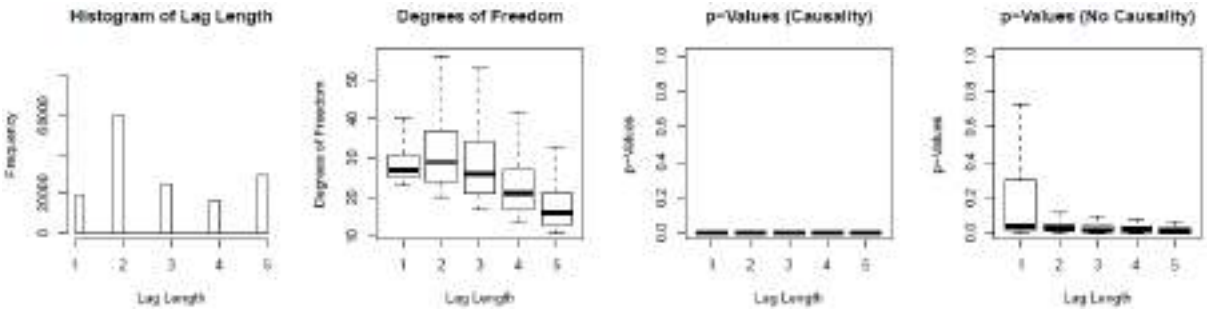
Appendix A2

Figure 7a: Prevalence of Overfitting Bias for DGP1a in the Presence of Theory Confirmation Bias ($h = 75$)



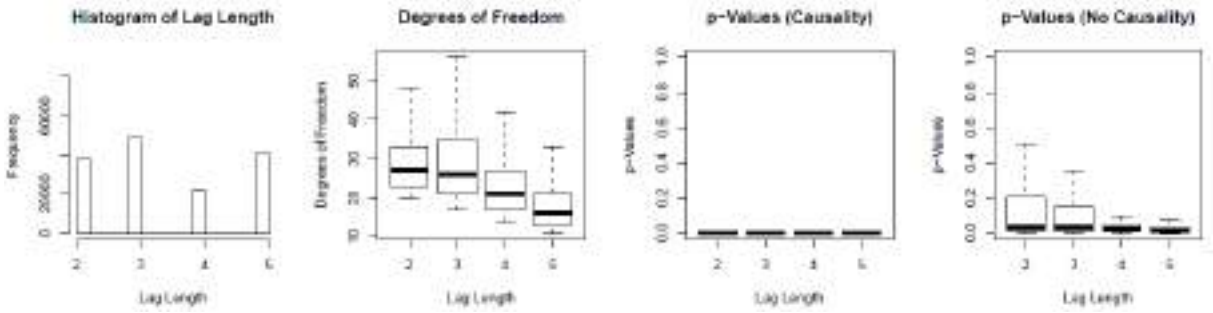
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) resulting in 150,000 observations for $\mu = 40$, $\sigma^2 = 100$, $\Omega = \mathrm{I}$, and $h = 75$. The second column presents 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.

Figure 7b: Prevalence of Overfitting Bias for DGP1b in the Presence of Theory Confirmation Bias ($h = 75$)



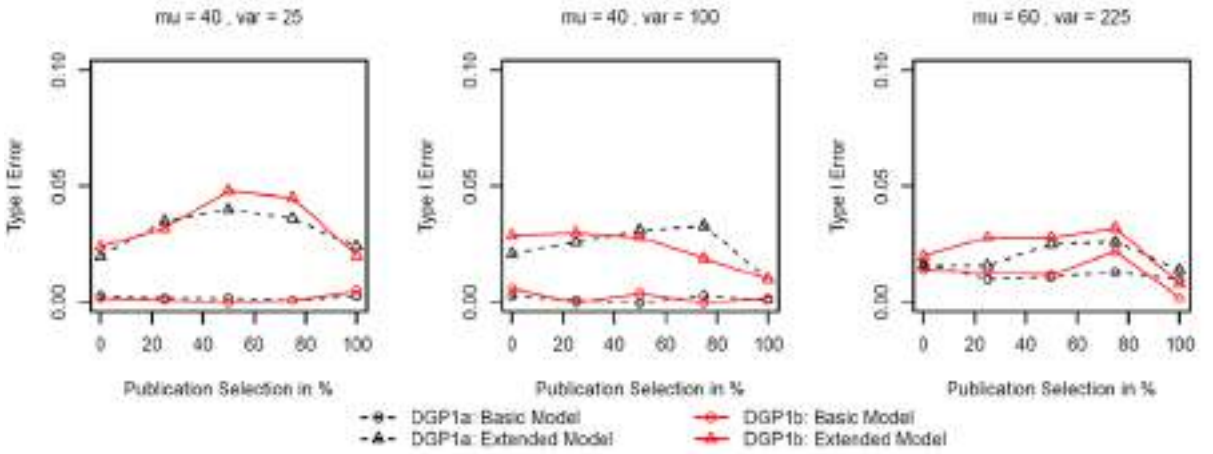
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) resulting in 150,000 observations for $\mu = 40$, $\sigma^2 = 100$, $\Omega = \mathrm{I}$, and $h = 75$. The second column presents 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.

Figure 7c: Prevalence of Overfitting Bias for DGP2b in the Presence of Theory Confirmation Bias ($h = 75$)



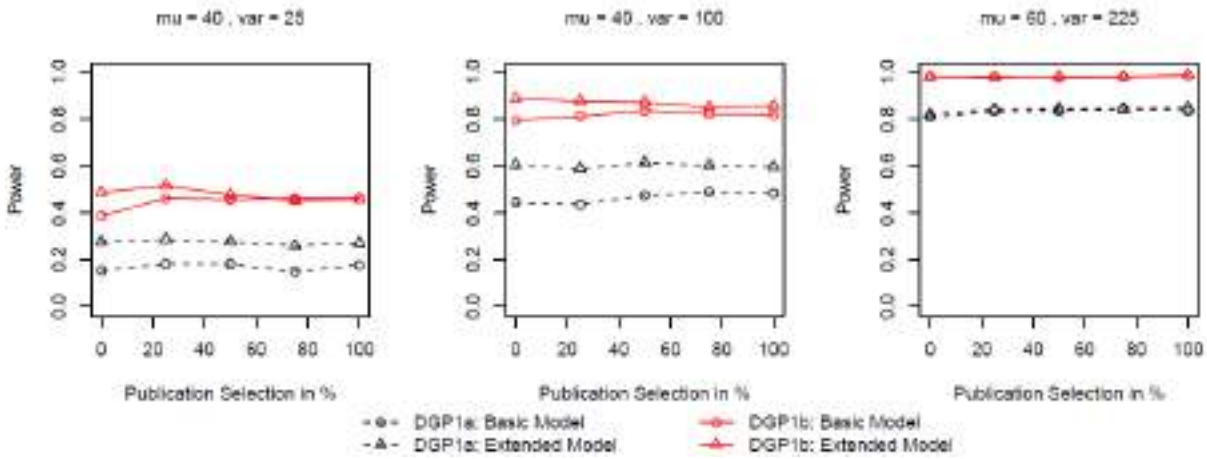
Notes: The first column shows the histograms of selected lag lengths in simulated primary studies for DGP2a across all meta-analysis sample sizes ($s = 10, 20, 40, 80$) resulting in 150,000 observations for $\mu = 40$, $\sigma^2 = 100$, $\Omega = \mathrm{I}$, and $h = 75$. The second column presents 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 0.01% primary studies and these findings are not reported.

Figure 8a: Type I Errors of Meta-Regression Models for DGP1a and DGP1b in the Presence of Theory-Confirmation Bias



Notes: Type I errors of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega = \mathrm{I}$ are reported as function of publication selection ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Figure 9a: Power of Meta-Regression Models for DGP1 in the Presence of Theory-Confirmation Bias



Notes: Power curves of $H_0: \beta_B^{gc} \leq 0$ (circles) and $H_0: \beta_E^{gc} \leq 0$ (triangles) for DGP1a (black) and DGP1b (red) with $\Omega = \mathrm{I}$ are reported as function of publication selection ($h = 0, 25, 50, 75, 100$) with $s = 40$ for small primary sample sizes distributions in column one and two and a larger primary sample size distribution in column three.

Appendix A3

Table 4: Studies Included in the Empirical Application

Authors and date	Countries	Control variables
Adom (2011)	GHA	-
Alam *et al*. (2011)	IND	Empl., capital, CO_2
Bowden and Payne (2009)	USA	Empl., capital
Ciarreta *et al*. (2009)	PRT	Energy Pr.
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	Empl., capital
Payne (2010)	USA	Empl., capital
Sari and Soytas (2009)	DZA, IND, NGA, SAU, VEN	Empl., CO_2
Soytas *et al*. (2007)	USA	Empl., capital, CO_2
Soytas and Sari (2009)	TUR	Empl., capital, CO_2
Vaona (2012)	ITA	-
Wolde-Rufael (2009)	17 African countries	Empl.; capital
Wolde-Rufael (2010a)	IND	Empl.; capital
Wolde-Rufael (2010b)	CHN; IND; JPN; KOR; ZAF; USA	Empl.; capital
Wolde-Rufael and Menyah (2010)	9 developed countries	Empl.; capital
Zachariadis (2007)	G7 countries	-
Zhang and Cheng (2009)	CHN	Capital; CO_2; popul.
Ziramba (2009)	ZAF	Empl.

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
	(0.98)	(1.22)	(3.42)	(1.11)	(1.64)	(4.95)
Df	-0.46*	-0.02	-0.18	-0.44*	-0.05	-0.43
	(0.19)	(0.18)	(0.63)	(0.20)	(0.24)	(0.87)
lags		0.76**	0.48		0.68*	0.57†
		(0.23)	(0.35)		(0.31)	(0.32)
KL			-0.11			-1.68
			(4.06)			(5.10)
KL*df			0.14			0.31
			(0.77)			(0.98)
Other			-1.84			-5.37
			(4.51)			(5.78)
Other*df			0.33			1.04
			(0.86)			(1.10)
Obs.	123	123	123	123	123	123
Adj. R^2	0.10	0.17	0.18	0.08	0.13	0.13

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Notes: Bootstrapped standard errors in parentheses. We bootstrap primary studies rather than single Granger causality tests to account for the dependence of multiple Granger causality tests per primary study. Significance codes represent a two-sided t-test. One-sided t-tests representing the test for a positive relation of z^{gc} and $\sqrt{df}$ are discussed in the text.