Universität Augsburg
Mathematisch-Naturwissenschaftlich-Technische Fakultät
Institut für Materials Resource Management
Prof. Dr. Andreas Rathgeber

Universität
Augsburg
University

Are speculators driving commodity market prices?
A quantitative literature review

MASTERARBEIT

zur Erlangung des akademischen Grades

„Master of Science“

Betreut von: Jerome Geyer-Klingeberg (MSc.)

Vorgelegt von: Martin Hillenbrand
1216326
Wirtschaftsingenieurwesen, WS 2019 / 2020
0157/52468404, hillenbrandmartin@posteo.de

Abgabetermin: 05.05.2020

Abstract

According to the current state of research, there is little supporting evidence that the recent increase in financial speculation has influenced commodity markets. In summary, this master thesis shows that little to no genuine overall speculation effect is detectable for soft commodity, energy and metal markets on the average of the investigated primary literature. A best practice model derived from multiple meta regression analysis supports the hypothesis that Granger non-causality between speculation and commodity markets cannot be rejected for a significance level of 5%. In total the literature sample of 46 empirical studies published between 1997 and 2018 analyzed the influence of financial speculation on commodity markets. The sample contains 1,560 Granger non-causality test results. The underlying literature has a large heterogeneity in their findings. Multiple meta regression analysis helps explaining this heterogeneity by including 28 moderation variables examining different study and test designs. The investigated commodity groups, used test setup, investigated time period and the ranking of the underlying studies are in large parts responsible for the variation of test results in the primary literature.

Table of content

List of figures

List of tables

List of abbreviations

ADL	Autoregressive distributed lag
AIC	Akaike information criterion
AJG	Academic Journal Guide
BIC	Bayesian information criterion
CFTC	Commodity Futures Trading Commission
DF	Degrees of freedom
EMH	Efficient market hypothesis
ESMA	European Securities and Markets Authority
FAT	Funnel asymmetry test
GC	Granger non-causality
G-to-S	General to specific
H2M	Hedging to market
IQR	Interquartile range
M2H	Market to hedging
M2S	Market to speculation
MIDFID II	Markets in Financial Instruments Directive 2
MRA	Meta-regression analysis
MSE	Mean squared error
NLP	Natural language processing
OI	Open interest
OTC	Over-the-counter
OLS	Ordinary least squares
PET	Precision-effect test
RePEc	Research Papers in Economics
S2M	Speculation to market
SE	Standard error
SJR	Scientific journal-ranking
VAR	Vector autoregressive regression
VEC	Vector error correction
VIF	Variance inflation factor
WLS	Weighted least squares

1 Introduction

The question if commodity speculation is harmful or not has been the subject of public debate, especially since the food product price increase during the years 2006 to 2008. Non-governmental organizations (NGOs) argue that index-focused investors increased commodity prices through their indirect holdings of long positions in commodity futures used by the providers of index funds for hedging. (Gilbert 2009, p. 1ff)

Speculation is often regarded as major cause for disturbances, like increasing price volatility and spillover effects from financial markets to commodity markets. In case of soft commodities, this speculative influence is often criticized to harm food security and the essential supply for people in the third world. The consequences of this debate are the halt of selling commodity related products by most European banks and strict regulations of commodity future trading. After the recent financial crisis, the European Union launched the second Markets in Financial Instruments Directive (MIFID II), ensuring that position limits for commodity derivatives, including all commodity contracts and all over-the-counter (OTC) positions, are implemented by the European Securities and Markets Authority (ESMA). In the US, the Dodd-Frank act calls for stronger limitations for agricultural commodities speculation. This resulted in the introduction of tighter position limits on exchange-traded contracts of 28 commodities by the Commodity Future Trading Commission (CFTC). (Massot et al. 2013, p. 1ff)

The clear and strong public and political opinion on the negative effect of speculation is rather surprising when compared to academic literature. There are numerous theoretical and statistical studies analyzing the impact of speculation on commodity futures markets. But there is no well-accepted agreement among economists if there is a positive or negative impact or an impact at all. Empirical findings are rather mixed, which may be because the studies differ strongly in terms of focus variables, commodities under investigation, the method for measurement of speculation, or the timeframe investigated. (Haase et al. 2016, p. 2ff)

The scientific and public debate is not new. Since the 19th century, speculation on financial markets in general and speculation on futures markets are in particular subjects of research and legislation (for example Fürst (1896)). Prominent economists like Kaldor (1976), Keynes (1930), Hicks (1946), Friedman (1953), Working (1953, 1961) and Telser (1959, 1967) have published research papers about speculation. Therefore, it is not surprising that there is a seemingly inexhaustible amount of studies on this topic. A few summaries and meta studies have been published, which have all analyzed different, particular aspects of literature. (Haase et al. 2016, p.1f)

The following is an overview of the core studies from this research field.

Haase et al. (2016) have analyzed a total of 100 published and unpublished empirical papers with six different focus variables (price levels, returns, risk premium, spreads, volatilities and spillover effects). They find strong evidence for a reinforcing effect of speculation on prices. For food commodities they find no evidence of a predominantly weakening or reinforcing influence of speculation.

Meijerink et al. (2012) reviewed a total of 40 quantitative and qualitative studies dealing with index speculation and agricultural commodities. They find no consistent evidence that large inflow of speculative capital, especially by index funds, has led to higher prices or more volatility in the mid and long term. They found certain evidence for very small and short-term volatility effects.

Will et al. (2012) surveyed 35 empirical studies addressing the impact of speculation on soft commodities. They found little evidence that the increase in financial speculation has caused an increase in price levels or volatility. Although some of the studies are criticizing the adverse effect of speculation.

Brümmer et al. (2013) use a more qualitative approach to classify and evaluate the literature. They conclude that the observed price volatility increases in recent times seem to be in broad agreement with fundamental factors. But they acknowledge that methodological difficulties and irregularities might distort the results.

All these studies have drawbacks, such as they used a qualitative, non-empirical approach or they were limited to a single commodity market.

Haase et al. (2016) for example deployed an adapted form of vote counting. Vote counting has some disadvantages, like the loss of information and the production of misleading results. (Stanley and Doucouliagos 2016, p. 43f)

Simple averaging of coefficients or test statistics across studies is imprecise, because effects like publication bias and misspecification biases distort the results (Bruns et al. 2014, p. 1). Publication bias is the tendency of authors and journals to publish mainly statistically significant or theory-conforming results (Card and Krueger 1995, p. 239f).

Sometimes studies strive to find significant results, for example through data mining or other techniques, even when there are no real effects in the underlying data. In this context Ioannidis (2005) claimed that, most published research findings are false. (Ioannidis 2005, p. 1ff)

In many areas of economics and finance Granger non-causality (GC) tests are applied (Ang 2008, p. 557ff). There are also a lot of studies analyzing the relation of speculation and

commodity markets using GC tests. But the results of GC testing are often fragile and unstable (Lee and Yang 2012, p. 393). In order to overcome this problem, meta-regression analysis (MRA), a sub-method of meta-analysis, is a suitable tool to summarize, integrate, correct and evaluate those research findings (Stanley and Doucouliagos 2016, p. 12).

In general meta-analysis is a method for aggregating results of individual empirical studies in order to increase statistical power and to remove confounding effects (Stanley 2001, p. 1f). MRA is a more effective meta-analysis approach to examine the impact of moderator variables on the results of studies using regression techniques (Thompson and Higgins 2002, p. 1559f).

In this master thesis an MRA approach is used to test 46 empirical studies using GC test techniques for genuine effects, explain heterogeneity and test for publication and misspecification biases.

2 Speculation theory and related literature

The following is a brief introduction to the theoretical background and different hypothesizes of speculation. Then the theory of the Granger non-causality test approach is presented.

2.1 Speculation theory

Over the past decade, the number of empirical studies analyzing the influence of financial speculation on commodity markets has increased dramatically. At the same time, the public debate on speculation rose, primarily led by NGOs, especially during and past the food product price increase in the years 2006 to 2008. (Gilbert 2009, p. 1ff)

Speculation is sometimes mixed-up with gambling, especially in the public opinion. But Keynes (1930) has already distinguished between gambling and speculation. He applies the term gambling to situations in which risk is not calculable or not following a normal distribution, such as the game of roulette. Speculation on the other hand is a situation in which the risk is calculable and normally distributed, such as life insurances. The dividing criterion is the amount of information available to the actor in both cases. Therefore, the possession of superior knowledge is the crucial distinction between speculators and gamblers. (Amato et al. 2012, p. 8)

Different theories and hypothesizes were developed on the question if speculation is influencing the commodity markets and if it has stabilizing or destabilizing character.

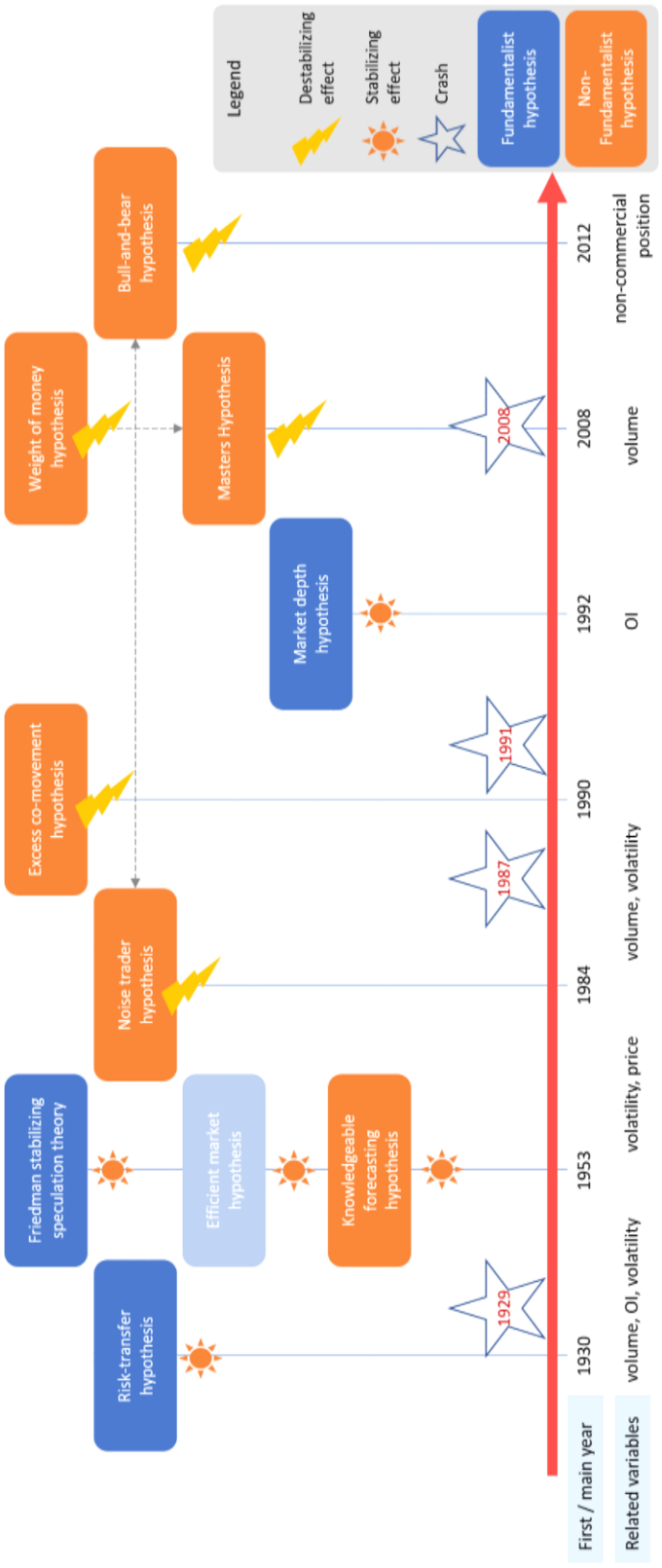

Figure 1: Fundamental and non-fundamental speculation hypothesizes

Source: Own source

There are two different types of hypothesizes. Figure 1 shows the development of the different hypothesizes over time.

The fundamental hypothesizes are mainly the theories of Keynes (1930) and Friedman (1953) together with the, non-directly speculation linked, efficient market hypothesis (Malkiel and Fama 1970, p. 1ff). They state that financial market prices are determined primarily by fundamental factors.

The non-fundamental hypothesizes on the other side argue that rational speculation based on fundamental information is not the only driver of speculation.

2.1.1 Friedman speculative stabilizing theory and Efficient market hypothesis

The traditional speculative stabilizing theory by Friedman (1953) argues that, if speculation is generally considered to be destabilizing, this is largely equivalent to saying that speculators lose money, since speculation can only be destabilizing if speculators sell when the price is low and buy when it is high. Friedman himself warns that this is a simplified generalization on a complex problem. It could be, that professional speculators might make money while many amateur speculators regularly lose larger sums.

Today Friedman's hypothesis is included in the more famous efficient market hypothesis (EMH). In simple terms it states that markets are generally efficient in processing instantaneously all available information about market fundamentals and macroeconomic development. Market participants are acting rational. And prices are randomly changing. As stated before, in consequence uninformed speculation cannot be profitable in a long term. That means commodity prices are not distorted in a systematic and/or persistent way by speculation. (Ederer et al. 2013, p. 9ff)

According to Friedman's hypothesis and the EMH financialization of markets and other speculative activity have no negative influence on the market but have a smoothing effect on the price formation and volatility instead. (Friedman 1953, p. 175; Steiner and Bruns 2007, p. 5; Buyuksahin and Harris 2011, p. 170f)

2.1.2 Risk-transfer hypothesis

Keynes (1930) and Hicks (1946) suggested that speculation fills a demand / offer gap in future markets. Sellers and buyers of goods, so called hedgers, want to reduce their price risk through the future market, but often do not seek for the exact same (symmetrical) kind of contracts. Speculators are relatively risk-tolerant and are rewarded for accepting the price risk from more risk-averse hedgers. This justifies the presence of speculators mediating between not perfectly symmetrical needs. (Hirshleifer 1977, p. 975; Amato et al. 2012, p. 31)

It should be considered that Keynes himself engaged in speculation and did not always act according to his idea of speculation. He was aware, that even if other speculators knew that prices were below their equilibrium level, they would wait to buy until "famine" arose and he would therefore wait as well. (Amato et al. 2012) This is an indication that not only fundamental hypothesizes exist.

2.1.3 Knowledgeable forecasting hypothesis

Working (1953, 1961) interprets what may look like risk-transfer behavior as the interaction of traders with more or less optimistic beliefs about the future affecting prices. A trader who expects prices to rise will make speculative purchases and one who expects them to fall will sell. In this sense the future markets provide an instrumentality whereby a consensus of beliefs about future supply-demand influences is brought to bear (by establishing a current price for later deliveries) upon current production-consumption decisions. (Hirshleifer 1977, p. 975)

However, this means that market behavior is no longer only based on fundamental information.

2.1.4 Noise trader hypothesis and bull-and-bear hypothesis

The "noise trader hypothesis" or "bull-and-bear hypothesis", suggest that in addition to fundamental factors, imperfect knowledge and the expectations of heterogeneous agents cause speculative activity which influences commodity prices. We can classify traders in roughly two categories: On the one hand, there are informed traders, who are basing their trading activities on fundamental factors and other dynamics in physical markets. These include commercial traders as well as some speculators classified as non-commercial traders, such as commodity trading advisors. On the other hand, there are uninformed traders, who may act upon beliefs or sentiments not based on fundamentals or employ technical trading strategies which are not directly based on fundamental information but on past price developments. The uninformed traders' strategies can lead to herding, which means, for example, the tendency of individual traders to copy actions of large groups instead of doing their own research and basing their decisions on their own information. Thus, this category of traders is also referred to as "noise traders". (Kyle 1985, p. 1)

This strategy, if dominant, may cause distortion of commodity market prices. Therefore, it might be rational for informed "fundamental" traders to follow the trend, even if unjustified by information about fundamentals. This can result in complex interrelations among different types of traders. (UNCTAD 2011, p. 18f; Ederer et al. 2013, p. 9)

Based on the noise-trader hypothesis, and in contrast to the efficient market hypothesis, higher financialization and therefore more speculative activity results in higher trading volume and rising market volatility.

2.1.5 Weight of money hypothesis and Masters hypothesis

Over the last couple of years commodity markets have been growing in volume extraordinarily. This was especially caused by big non-commercial traders, like index investors. That's why it's also called "financialization" (Pradhananga 2014, p. 3f). Those investors generally take very large positions on one side of the market. Those individual market participants may make position changes that are so large relative to the size of commodity markets that they move prices temporarily or even long term. (Ederer et al. 2013, p. 9f; UNCTAD 2011, p. 18ff)

The hypothesis is also called "Masters Hypothesis" due to the statement of Mr. Masters, a former investor, who claimed to the U.S. Congress and the Commodity Futures Trading Commission (CFTC), that large long-only investment was a major driver of the 2006 to 2008 spike in commodity futures prices and energy futures prices (Etienne et al. 2017, p. 52; Irwin and Sanders 2012, p. 3). Based on the "weight of money hypothesis" more speculation results in higher trading volume and more open interest (OI) on the long side.

2.1.6 Excess co-movement hypothesis

Co-movement of commodity prices can often be explained by a fundamental supply and demand relationship or common macroeconomic effects. This includes, that the interest rate level plays an important role in the formation of future prices. (Hull 2006, p. 145ff)

The question remains, if co-movement can be exclusively explained by those factors. If not, it is called excess co-movement and it might be explained by speculation. Financial investors and speculators may base their trading of different commodities and other asset classes on non-fundamental information. Unlike commercial traders they also use strategies of technical trading that are used by noise or uninformed traders and may produce co-movement without fundamental information. (Ederer et al. 2013, p. 10; Pindyck and Rotemberg 1988, p. 1ff)

Results from Deb et al. (1996) for an observed time span from 1960 until 1992 only obtain weak evidence for excess co-movement (ECM). This is not surprising since it is assumed that ECM occurs in the nearer past. Lescaroux (2009) analyzed the years 1980 to 2008 and found no strong links between commodity prices, when filtering out the influence of supply and demand. Natanelov et al. (2011) on the other hand identified strong linkages between soft commodities and crude oil prices for the sample period from 1993 to 2007 and

suggested a dynamic concept and understanding of co-movement which is highly influenced by economic and policy development.

2.1.7 Spillover hypothesis

There are different theories on the influence of future trading and speculation activity on spot markets. All hypothesizes influencing the future markets, especially those mentioned before, are assumed to also influence spot markets because of the physical interlink between spot and future markets. Some researchers argue that the pure existence of future trading can increase spot volatility (Nath and Lingareddy 2008, p. 23). Research findings of Figlewski (1981) indicate that future markets could attract new hedging without attracting enough speculation to permit effective risk transfer. The pressure in the future market could spill over to spot markets where dealers and market makers end up bearing the risks from both the spot and the future market. Yang et al. (2005) identify unexpected futures trading volume to uni-directionally cause spot price volatility for most of the commodities. In context of the Indian commodities market, research findings of Nath and Lingareddy (2008) state that the futures market has not helped in reducing cyclical/seasonal fluctuations in the spot market, which was one of the benefits of future trading proposed by Friedman and Keynes (cf. chapter 2.1.1 and 2.1.2).

Furthermore future trading has led to an increase in volatility in the spot market for some commodities. (Gupta et al. 2018, p. 54)

2.1.8 Market depth hypothesis

Market depth is another variable of future trading, which has been defined by Kyle (1985, p. 1330) as " … the ability of the market to absorb quantities without having a large effect on price". Bessembinder and Seguin (1992) argue that the introduction of future trading improves the liquidity and market depth due to the existence of market makers. Consequently, the transmission of information between future and spot markets improves. Higher market depth also reduces spot price volatility. They suggest the use of OI to measure market depth. (Bessembinder and Seguin 1992, p. 1ff)

Due to the limitation of this master thesis to Granger non-causality tests and to gain reliable and comparable results not all the presented hypothesizes are included in the later analysis. But they are still listed here to give a complete overview of the speculation theory.

2.2 Granger non-causality theory

For the understanding of the following meta-analysis of Granger non-causality test data it is important to first understand what Granger non-causality means.[1] The concept of causality developed by Granger (1969) is fairly easy and has become quite popular in recent years. One of the basic, underlying conditions is that a cause cannot come after the effect. If a variable *x* affects a variable *y*, the former should help improving the predictions of the later variable. To formalize this idea, suppose that Ω_t is the information set containing all the relevant information in the universe available up to and including period *t*. And $y_t(h|\Omega_t)$ is the optimal minimum mean squared error (MSE) h-step predictor of the process y_t at origin *t*, based on the information in Ω_t.The corresponding forecast MSE will be denoted by $\sum_y(h|\Omega_t)$. (Lütkepohl 2007, p. 41ff)

The process x_t is said to cause y_t in Granger's sense if

Eq. 2.1

$$\sum_y(h|\Omega_t) < \sum_y(h|\Omega_t \setminus \{x_s|s \leq t)$$

For at least one $h = 1,2, \ldots$

$\Omega_t \setminus \{x_s|s \leq t)$ is the set information containing all the relevant information in the universe except for the information in the past and present of the x_t process. If one gets a more efficient prediction of y_t if the information in the x_t process is taken into account in addition to all the other information in the universe then x_t is Granger-causal for y_t. If x_t is not Granger causal for y_t there should be no difference between both parts of the equation 2.1. This can be tested by the null hypothesis in equation 2.2. (Stanley and Doucouliagos 2016, p. 35; Lütkepohl 2007, p. 41ff)

Eq. 2.2

$$\sum_y(h|\Omega_t) \neq \sum_y(h|\Omega_t \setminus \{x_s|s \leq t)$$

Usually all the relevant information in the universe (Ω_t) is not available to be used in applied econometrics and therefore it is not possible to deployed equation 2.1 or 2.2

[1] In the literature the term Granger causality test is often used instead of Granger non-causality test. Both describe the same, and in this thesis the term "Granger non-causality test" or the abbreviation GC are used.

directly. Therefore in practice, Granger non-causality tests are usually based on improved linear predictions within a specific model. (Lütkepohl 2007, p. 41ff)

2.2.1 Linear Granger non-causality test

One more practical and one of the most common approaches uses an autoregressive distributed lag (ADL) model, described by equation 2.3:

Eq. 2.3

$$y_t = \alpha + \sum_{i=1}^{n} \gamma_i \, y_{t-i} + \sum_{j=1}^{m} \beta_j \, x_{t-j} + error_{2,\,t}$$

y_t is the dependent and x_t the independent variable. The lags for variable *x* and *y* are *m* and *n*. The null hypothesis test $H_0: \beta_j = 0 \; \forall j = 1, \ldots, m$ is a F- or Chi²- test for Granger non-causality. (Sanders and Irwin 2011a, p. 43ff; Granger 1969, p. 427ff)

To perform a correct Granger non-causality test, it is essential to have information about the properties of the time series under consideration. For integrated time series a Wald test statistic for Granger non-causality in a VAR in levels follows non-standard asymptotic distributions and depends on nuisance parameters (Stock et al. 1990, p. 124; Toda and Phillips 1993, p. 1367ff). Testing for integration can be done by a variety of unit root tests, but all these tests suffer from low power in small samples. For first order integrated (I (1)) time series without cointegration, a Wald test in a VAR in first differences framework can be used for testing Granger non-causality. If the time series is first order integrated (I (1)) and cointegrated, a vector error correction (VEC) model is the appropriate testing framework. For not integrated (I (0)) time series, Granger non-causality can be directly tested using a VAR in levels as the unrestricted model. (Bruns and Stern 2015, p. 5f)

Testing for the order of integration and cointegration to decide which Granger non-causality testing framework should be used, can introduce pre-testing biases. (Bruns and Stern 2015, p. 5f)

2.2.2 Robust linear Granger non-causality test

Stationary time series variables are one of the prerequisites in the classical Granger model. The presence of a unit root would change all asymptotic properties of estimators. In many cases guaranteeing that all considered time series are stationary is hard to impossible, because the power of standard tests for a unit root is very low. In case of a unit root in the considered times series, there is no asymptotic Chi² distribution under the null hypothesis for the standard test statistics like the F-test, the Wald test or the LR test. This means no Granger non-causality test by using the VAR model can be conducted if some variables have a unit root. This problem can be avoided by using first-difference series and the causal

relationship can be analyzed, but with more limited information than the level series. (Ding et al. 2014, p. 178f)

The robust Granger non-causality test is a remedy for this problem and it can be applied to avoid any pre-testing biases. Dolado and Lütkepohl (1996) developed a GC test, which is robust to the integration and cointegration properties of data. Toda and Yamamoto (1995) show that a VAR in levels can be augmented by a number of lags equal to the highest degree of integration d_{max}. Then, a (k+d_{max})th VAR order is calculated and the coefficients of the last lagged d_{max} vector are ignored.

Granger non-causality can therefore be tested by the following VAR model:

Eq. 2.4

$$Y_t = \alpha_0 + \sum_{i=1}^{k} \alpha_{1i}\, y_{t-i} + \sum_{j=k+1}^{d_{max}} \alpha_{2j}\, y_{t-j} + \sum_{i=1}^{k} \delta_{1i}\, x_{t-i} + \sum_{j=k+1}^{d_{max}} \delta_{2j}\, x_{t-j} + error_{1t}$$

From Eq. 2.4 causality in Granger-sense from x_t to y_t implies $\delta_{1i} \neq 0\ \forall_i$

Toda and Phillips (1993) and Toda and Yamamoto (1995) have developed a robust inference method, in which the standard test statistics in a VAR framework follow an asymptotic Chi² distribution under the null hypothesis even when variables are integrated or cointegrated in arbitrary order. (Ding et al. 2014, p. 178; Toda and Yamamoto 1995, p. 228ff)

2.2.3 Granger non-causality test in quantiles

In addition to the linear Granger causality tests there exist other tests. GC in quantiles is one of them. Chuang et al. (2009) suggest a Granger non-causality test in quantiles by using the quantile regression method introduced by Koenker and Bassett (1978) and the sup-Wald test by Koenker and Machado (1999). Several other parametric (e.g. Troster 2018), non-parametric (e.g. Jeong et al. 2012) and linear (e.g. Song and Taamouti 2020) in quantile GC test approaches exist. (Chuang et al. 2009, p. 1ff)

2.2.4 Non-parametric Granger non-causality test

The conventional approach of linear Granger causality has limited power compared to certain nonlinear alternatives, because it is based on the restrictive linear functional form (Naderian and Javan 2017, p. 30). Baek and Brock (1992) suggest a non-parametric, statistical technique for uncovering nonlinear causal relationships that cannot be discovered by standard linear causality tests. They assume that the variables are mutually independent and identically distributed. Hiemstra and Jones (1994) modify this test, so that it can also be applied to variables with short-term temporal dependences. Unlike in linear Granger non-causality tests, no methods have been developed in the literature to select optimal lag

values (Fujihara and Mougou 1997, p. 400f). Hiemstra and Jones (1994) proposed a non-parametric test for general (linear and non-linear) Granger non-causality. But this former commonly used test suffers from lack of power and can severely over-reject if the null hypothesis is true. (Diks and Panchenko 2005, p. 1)

Hence results using the Baek and Brock (1992) or the Hiemstra and Jones (1994) approach should be omitted. Diks and Panchenko (2006) have improved the approach and ensured that no more spurious results are produced. Other, not so frequently used non-parametric GC tests in applied economics and finance are proposed by Bell et al. (1996) and by Su and White (2008). (Diks and Panchenko 2006, p. 1648)

3 The meta-analysis data set

The goal of this meta study is to assess the findings of empirical studies which analyze the impact of speculation on commodity markets via Granger non-causality tests. Stanley et al. (2013) suggest reporting guidelines for meta-analysis of economics research and this thesis follows their suggestions. The range of papers, with 70 empirical studies, is smaller than other meta studies like Haase et al. (2016), Meijerink et al. (2012) and Will et al. (2012). Similar to the meta study of Haase et al. (2016) it is not restricted to agricultural commodities, but includes all commodities reported in the underlying study sample. The primary selection of empirical studies was made by the Chair of Finance- & Information Management of the Institute of Materials Resource Management at University Augsburg.

To ensure having a representative sample of empirical studies an additional systematic database search (Google Scholar) was conducted. Keywords were extracted from the primary set of studies with a Natural Language Processing (NLP) approach. Due to time limitations of this master thesis and difficulties with the processing this approach was discarded. No extended coding of additional empirical studies was conducted, but the primary sample of studies is expected to be a good representation of the whole empirical literature on this topic. Certain criteria must be fulfilled by primary articles under investigation to get consistent and comparable results in the meta-analysis. Some of the studies therefore had to be excluded. The reasons for excluding studies are documented in appendix B.

3.1 Design of empirical analysis

Studies must report results from Granger non-causality tests between a variable measuring financial speculation and one measuring commodity markets behavior. Therefore, the underlying Granger non-causality test for an influence of speculation on commodity

markets must indicate an independent variable *x*, for a measurement of speculation, and a dependent variable *y*, measuring the commodity market.

3.2 Reported results

Studies must provide statistical information required for MRA. In case of a standard MRA this includes regression coefficients, sample size and a precision measure of the regression estimate (such as standard errors). But GC tests normally don't report standard errors and the regression coefficients are not of interest, because the test statistics itself shall be analyzed. Therefore, F-statistics, z-, t- and Chi^2-statistics and/or p-values must be reported in the studies to identify results of the GC tests. Sample size and the number of lags are needed to calculate the degrees of freedom (DF) to analyze the impact of publication selection bias. The number of lags for the *x* variable is also needed to identify the possible impact of overfitting bias on the test results. Where possible (and necessary) information were calculated or assumed from data given in the studies. The calculations and operations are mainly done in R. The R code can be downloaded at the address indicated in appendix F. All operations not done in R are listed in appendix A. Assumptions and the description of the coding process is listed in appendix E. To ensure comparability, percentage values were transformed into decimal values.

3.3 Proxy variables and data

As stated before Granger non-causality tests control for the influence of *x* variables on *y* variables. Nevertheless there is no perfect variable for the identification of speculative activity. Alquist and Gervais (2013) and Buyuksahin and Harris (2011) present a variety of speculation measures, but point out that it remains challenging to identify effects from speculation.

Proxy variables that represent speculation and market behavior are used to observe the effect of speculation on commodity markets. The empirical literature suggests a wide range of speculation and market determinants. Table 1 describes the set of selected and grouped proxy indicators as well as non-assignable determinants analyzed in the MRA. The variable definitions are derived from the primary studies.

The full list of studies included is available in appendix B.

Table 1: Proxy variables for speculation and market behavior and their usage as x and y variables in Granger non-causality tests

Proxy variable groups	Definition	Included variables	x	y
return	First or more differenced price data.	Return (differenced price data), roll return, fund returns, variance of return etc.	2,179	2,425
volatility	Volatility of commodity prices.	Volatility, implied volatility, conditional volatility	1,197	1,247
price	Price data (in levels)	Price, price shock, etc.	274	470
other	Subsumed market variables	Spread, Liquidity, etc.	90	127
market behavior proxy variable group (sum)			3,740	4,269
position of speculators	Position data of non-commercial trader	Position of hedge funds, money manager, etc.	1,625	1,163
open interest	Open interest	Open interest of commodity index trader, etc.	79	70
volume	Trading volume	Trading volume, etc.	104	104
other	Subsumed speculation variables	Variance growth, etc.	22	18
speculation proxy variable group (sum)			1,830	1,355
Commercial / hedger position data	Position data of commercial trader	Position of manufacturers, etc.	618	598
hedging proxy variable group (sum)			618	598
Other position data	Non-assignable positions data	Position of all, position of non-reportable, etc.	165	135
other	Non-assignable other data	Fund rolling	4	0
other variable group (sum)			169	135
total proxy variables			**6357**	**6357**

3.4 Paper quality

The broad range of papers makes it necessary to apply a quality assessment of the papers. Stanley and Doucouliagos (2016) recommend collecting information about study quality while coding. They also recommend using the estimate's precision as indicator of quality. Therefore, in a normal MRA the standard error is coded. In this thesis, however, another approach, which is explained in chapter 4, is applied and, thus, the square root of degrees of freedom are used as precision estimate. If the degrees of freedom are not specified in the underlying study, they are calculated from the sample size and the lags.

In addition, several study quality classifiers were used. The following variables are coded to identify the papers' quality:

i. Number of Google citations[2]
ii. Research Papers in Economics score (RePEc)[3]
iii. Elsevier based scientific journal-ranking criterion for published papers for the year 2018 (SJR)[4]
iv. Academic journal guide score for the year 2018 (AJG)[5]

Google citations are the indicated citations of an individual scientific studies identified by Google Scholar in other studies. A higher citation number is probably an indication of a more relevant paper in the scientific community. (Gusenbauer 2019, p. 7ff)

The RePEc score is a journal scoring mainly based on download count and indicates the impact of a journal in the field of economics research (Zimmermann et al. 2020).

The SJR score for 2018 is a publicly available portal that includes journal indicators developed from the Scopus database. It covers most of the journals in the area of finance and economics research.

The fourth quality classifier is the AJG score for 2018 measuring the range and quality of journals for business and management research.

This thesis deploys four study quality factor, because one alone is often prone to errors and the databases are often incomplete (Gusenbauer 2019, p. 1). To obtain a complete and realistic assessment and to keep the number of moderation variables in the later MRA low, those four variables are combined to a self-developed paper ranking factor in equation 3.1.

[2] https://scholar.google.com/
[3] https://ideas.repec.org/
[4] https://www.scimagojr.com/
[5] https://charteredabs.org/academic-journal-guide-2018-view/

Eq. 3.1

$$ranking_i = \frac{\sum_j \frac{x_{i,j}}{\sum_{i=1}^{n} x_{i,j}}}{4}$$

$x = value\ of\ study\ quality\ classifier\ j$

$j = \{sjr2018, ajg2018, repec, googlecits\}$;

$i = individual\ study$;

$n = total\ number\ of\ studies$

This ranking factor is expected to account for quality differences between the studies not captured by the other variables. It can be used in the further multiple MRA explained in chapter 6.4.

3.5 Unpublished literature

Out of 117 authors, 83 of which the contact details could be found, mainly on Research Gate, were contacted and asked if they have omitted and/or unpublished studies related to GC tests for speculation in commodity markets. Out of 11 responses, only 2 had further studies. Those studies had either already been included in the data set, did not comply with the required criteria, or the data from these unpublished studies had been published in other studies that were already included in the data set.

4 Methodological approach

Meta-regression analysis is also described as "the regression analysis of regression analysis" (Stanley and Jarrell 2005, p. 1). The idea is, inspired by the "normal" meta-analysis, to aggregate the effect sizes of regression models.

One of the most important advantages of MRA over primary studies is that it minimizes random estimation error by averaging across the entire research record. Therefore, it allows to detect and correct for publication bias and other biases. The multiple MRA identifies variables explaining the systematic differences and variation of estimates in existing research results, also called heterogeneity.

There are two main steps in performing an MRA:

1) Synthesize the effect sizes from the primary studies
2) Explain the heterogeneity among the effect sizes by identifying study characteristics explaining this variation.

By pooling estimates across different studies, MRA minimizes estimation errors and allows inferences without depending on specific sample characteristics.

MRA can objectively control for any study-specific aspect, such as the model specification, variable definitions and other author-specific dimensions that might have an impact on the heterogenous findings, whereas the results of primary studies are subject to their specific research design. MRA adds value as it explicitly considers factors that are constant within a study but vary across studies, i.e. publication status, number of citations or commodity under examination. (Stanley and Doucouliagos 2016, p.3ff)

4.1 Choice of effect size

The main criterion for the selection of the effect size is that the estimates must be comparable within and across studies. (Stanley and Jarrell 2005, p. 301)

In this paper, the measure to be aggregated is the p value. As observable from Table 1, studies differ in the way they operationalize the proxy variables. Accordingly, the interpretation of the p-value heavily depends on the definition of the variables and other moderating factors.

The data required for MRA is not always directly reported in each study. For example, the degree of freedom is rarely found in the primary studies. Thus, they were inferred from the number of observations and the number of independent variables included. If p-values were not available, they were derived from the reported F- / z- / t- / Chi^2- values and the degree of freedom and lags. If the number of observations was not available, it was calculated from the start and end date of the sample and the frequency of the data. This calculation is explained in detail in appendix A.

4.2 Basic MRA model

An important challenge for every meta-analysis arises from publication selection bias, which appears when researchers and journals neglect results that are either statistically insignificant or inconsistent with theoretical predictions (Geyer-Klingeberg et al. 2019, p. 207). Without such a preference for certain outcomes, the estimated effect size β, in econometrics often a regression coefficient, should have the same expected value across different studies irrespective of their degrees of freedom (Bruns et al. 2014, p. 103).

Published studies require methodological innovation, which suppresses research using well-established methods, and research replicating and verifying prior studies. In addition, there may be a bias towards papers with significant test results in published work. (Haase et al. 2016, p.5)

Some researchers discard results from a publication if they do not comply with their preference for a certain study outcome. Reviewers and editors may be predisposed to accept papers consistent with the conventional view. And as a third cause for publication selection bias in general people may possess a predisposition to treat "statistically significant" results more favorably. Especially researchers with small samples and low precision will be forced to search more intensely across model specification, data and econometric techniques until they find larger estimates, in order to get statistically significant results. On the other side researchers with larger studies need not to search so hard from the practically infinite model specifications to find statistical significance and will be satisfied with smaller estimated effects. (Stanley and Doucouliagos 2016, p. 51ff)

Therefore, studies with higher degrees of freedom are expected to have less publication bias. Previous studies reveal that publication bias is a problem in many different areas of economics and other fields of science.

For a first investigation of publication bias, the standard FAT-PET meta-regression model is commonly used in economics. This model had to be modified for a meta-analysis of Granger non-causality test statistics. (Bruns et al. 2014, p. 2)

When there exists publication bias the reported effect size, in econometrics typically a regression coefficient, is positively correlated with its standard error. Without a publication selection bias, estimates and their standard errors will be independent, as required by the conventional t-test and guaranteed by random sampling theory. Therefor the magnitude of the reported estimate will depend on its standard error shown in Equation (1).

Eq. 4.1

$$effect_{ij} = \beta_0 + \beta_1 SE_i + \varepsilon_i$$

$effect_{ij}$ indicates an individual estimate of Granger non-causality test *i* of study *j* and SE_{ij} is its standard error. $\beta_1 SE_{ij}$ is a model for publication selection bias and β_0 is a correction for publication bias. ε_{ij} is the error term, which is not expected to be independently and identically distributed. The variance of $effect_{ij}$ and therefore ε_{ij} will typically vary from one estimate to the next. Thus, equation 4.1 has heteroscedasticity and should never be estimated by ordinary least squares (OLS) but by weighted least squares (WLS). Hence the WLS version of Eq. 4.1 can be calculated dividing equation 4.1 through by SE_{ij}. (Bruns et al. 2014, p. 103ff)

Eq. 4.2

$$t_{ij} = \beta_1 + \beta_0 \left(\frac{1}{SE_{ij}} \right) + \upsilon_{ij}$$

Where t_{ij} is the t-statistic of each individual estimated empirical effect, $1/SE_{ij}$ is its precision and $\upsilon_{ij} = \varepsilon_{ij}/SE_{ij}$, which should make its variance approximately constant. Without publication and misspecification biases and abstracting from genuine heterogeneity, the estimated effect size should have the same expected value across different studies independently of their degrees of freedom. The precision of $1/SE_{ij}$ is impractical for analyzing the results of GC tests, due to the fact, that the used test statistics do not have associated standard errors. Therefore, the inverse of the standard error is replaced with the square root of the degrees of freedom of the regressions in the underlying studies, which results in equation (4.3). Stanley and Doucouliagos (2016) suggest using the variance instead of the standard error for equation 4.1, but because of the replacement of the precision factor this is not possible. (Bruns et al. 2014, p. 103ff)

Eq. 4.3

$$t_{ij} = \beta_1 + \beta_0 \left(DF_{ij}^{0.5} \right) + \varpi_{ij}$$

Where $DF_{ij}^{0.5}$ refers to the square root of the degrees of freedom of the corresponding GC test *i* of study *j*.

The standard restriction test statistics of Granger non-causality tests have a F- or Chi²-distribution. These must be converted to statistics with a common distribution with properties that are suitable for the meta-regression analysis. This can be achieved by transforming the p-values of the original test statistics to standard normal variates using the probit function, like in equation (4.4). This ensures that the p-values are standard normally distributed under the null of Granger noncausality resulting in required residual properties. For meta-regression analysis the standard normal distribution has an additional advantage over the commonly used t-distribution because it is not affected by the degrees of freedom. (Bruns et al. 2014, p. 103ff)

Eq. 4.4

$$-probit\left(p_{ij} \right) = \beta_1 + \beta_0 \left(DF_{ij}^{0.5} \right) + \varpi_{ij}$$

The probit-values are multiplied with -1 so that higher values indicate a lower p-value and the results are easier to understand and interpret. In the following the expression probit-value for negative probit is used, if not stated otherwise. (Bruns et al. 2014, p. 104f)

The test of $H_0: \beta_1 = 0$, also called funnel asymmetry test (FAT) serves as a test whether there is publication selection or not.

The test of $H_0: \beta_0 = 0$, on the other hand is the so called precision-effect test (PET), which is used to test for genuine underlying empirical effect beyond the potential distortion due to publication selection. (Stanley and Doucouliagos 2016, p. 60ff)

Both tests, however useful they may be, have their weaknesses. The FAT is known to have low power. (Stanley and Doucouliagos 2016, p. 64ff) While the PET is usually powerful enough, it can have inflated type 1 errors and therefor mistakenly detect effects that are not there. (Stanley 2007, p. 114)

These inflated type 1 errors occur when there is much excess unexplained heterogeneity in the meta-regression model. When there is strong evidence ($p<0.001$) that the majority of the MRA error variance is due to unexplained heterogeneity (reject $H_0: \sigma_e^2<2$), a multiple MRA should be rather used to explain the systematic heterogeneity. (Stanley and Doucouliagos 2016, p. 64f)

4.3 Augmented MRA model

The basic MRA model obtained from Bruns and Stern (2015) and Stanley et al. (2013) can be further extended to account for other biases, like the overfitting bias.

Granger non-causality test statistics based on VAR models are very sensitive to the chosen lag length. Normally the true lag length is unknown and objective criteria are used to select a lag length. Choosing the lag length in a VAR model is mostly an empirical question, as economic theory is usually not very specific about the temporal dimension of economic dynamics. Among others the Akaike information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (Schwarz 1978), also known as the Schwarz information criterion, are the most commonly ones. But those information criteria are known to have a tendency of over- or underestimating the true lag length. (Bruns and Stern 2015, p. 2f)

Overfitting bias is more prevalent for AIC and underfitting bias for BIC (Bruns and Stern 2015, p. 21; Lütkepohl 2007, p. 150f). The resulting overfitted and underfitted models tend to lead to over-rejection and under-rejection of the Granger non-causality tests compared to models using the true lag length (Zapata and Rambaldi 1997, p. 2ff). Similar to publication selection the overfitting must not always be done intentionally by the author but can also happen unconsciously by searching for statistically significant results (Bruns and Stern 2015, p. 3f). Especially small samples are vulnerable to over- and underfitting bias (Bruns and Stern 2015, p. 2f; Gonzalo and Pitarakis 2002, p. 406ff)

To control for overfitting biases, the MRA model 4.4 has to be adjusted to account for the lag structure of the independent variable. Bruns and Stern (2015) have developed an MRA model Eq. 4.5 testing for genuine Granger non-causality while taking both, overfitting and sampling errors, into account.

Eq. 4.5

$$-probit(p_{ij}) = \beta_1 + \beta_0(DF_{ij}^{0.5}) + \beta_2 lags_{ij} + \varpi_{ij}$$

Bruns and Stern (2019) argue that even when information criteria are used for the selection of the lag length overfitting might occur, especially in small samples. To further investigate the influence of the chosen information criterion on the reported results, a moderator variable controlling for the usage of AIC is employed in the multiple MRA approach.

4.4 Multiple MRA model

The standard test for heterogeneity is Cochran's Q-test. But there is a statistical problem with using the Q-test for heterogeneity testing. It is widely known to have low power, thus finding no heterogeneity may only reflect the limitations of the test rather than the true homogeneity of the research record. Stanley and Doucouliagos (2016) suggest abandoning the Q-test altogether and just proceed as if there is heterogeneity in all cases. They justify this advice by their experience that across several dozen meta-analyses of economics research, the Q-test always indicates heterogeneity, despite its low power. Thus, they say it is unlikely to matter in practice whether or not the Q-test is calculated. (Stanley and Doucouliagos 2016, p. 48f)

Therefore, regardless of the outcome of a potential Q-test, heterogeneity is assumed and the following multiple MRA method is used to identify potential heterogeneity.

The MRA model in equation 4.5 assumes that differences across studies arise from random sampling error, publication, and overfitting bias. But the data set is obtained from different studies using different methodological approaches and different underlying data, leading to heterogeneity. To account for this heterogeneity the MRA model is extended for moderator variables, which are suspected to drive the variation in primary test results. This leads to the multiple MRA model in equation 4.6.

Eq. 4.6

$$-probit(p_{ij}) = \beta_1 + \beta_0(DF_{ij}^{0.5}) + \beta_2 lags_{ij} + \sum_{m=1}^{M} \beta_m Z_{mij} + \varpi_{ij}$$

Here Z_{mij} is the moderator variable for moderator m of Granger non-causality test i in study j. The moderator variables are presented in Table 2. (Stanley et al. 2013, p. 100f)

4.5 MRA model specification and testing

To be able to conduct a linear regression certain assumption must be valid. To ensure the validity and quality of the used linear regression models, they are tested for non-linearity, correlation of error terms, non-constant variance of error terms, outliers, high-leverage points, and (multi-) collinearity. (James et al. 2013, p. 92)

4.5.1 Non-linearity

The linear regression model assumes that there is a linear relationship between the predictors and the response. If this is not the case, then all conclusions drawn from the fit are questionable and the prediction accuracy of the model can be significantly reduced. (James et al. 2013, p. 92f)

4.5.2 Correlations

If estimates or error terms in an MRA model are correlated, this might falsify the results. Especially correlation of error terms might produce an unwarranted sense of confidence producing false-positive results. In meta regression analysis it is highly probable that certain correlations in the data exist. For example, studies can report multiple results, authors may have written several studies on the same sample, studies may have used the same database or method etc. To fight this problem several approaches can be used. Correlations like within-study correlation and across-study correlation based on author-level can be observed easily. Using clustered standard errors can help solving this problem. (James et al. 2013, p. 93f)

More conventional autocorrelation is also possible in MRA, but more likely to be seen in macroeconomics (Stanley and Doucouliagos 2016, p. 68). Autocorrelation can be identified by several tests or plotting the residuals and searching for a "tracking" pattern. "Tracking" occurs, if the error terms are positively correlated, and as a result adjacent residual may have similar values. On the other hand if errors are uncorrelated, there should be no discernible pattern. (James et al. 2013, p. 93f)

With 56 studies and 6357 collected p-values (all subsets included), the median of observations per study is 504. As there are multiple estimates per study within-study dependencies exist, which violate the assumption that disturbance terms are independently and identically distributed (Greene 2003, p. 250ff; Geyer-Klingeberg et al. 2019, p. 207)

Some authors have written more than one study on the same sample, therefor across-study dependencies on an author level exist. Clustered standard errors are used with clusters on the individual study- and author-level to correct for those dependencies. The used two-

dimension approach follows Wooldridge (2003), Cameron et al. (2011) and Petersen (2009) and the references therein. (Arai 2015, p. 1ff)

Stanley and Doucouliagos (2016) suggested this remedy for within and across study dependency in MRA research.

4.5.3 Heteroscedasticity

Non-constant variance of error terms, also called heteroscedasticity, can distort the results of a linear regression. This variance or heteroscedasticity can occur because each Granger non-causality test is based on different primary studies, with different sample sizes and different test characteristics. A studentized Breusch-Pagan test, like suggested by Koenker (1981), reveals potential heteroscedasticity. (James et al. 2013, p. 95ff)

The usage of weighted least squares (WLS) instead of ordinary least squares (OLS) is suggested by Stanley and Doucouliagos (2016) as a way to reduce heteroscedasticity.

Three different approaches were used as weights.

1) WLS 1: The square root of degrees of freedom serves as precision weighting factor (precision weight). This ensures that more "precise" test results receive a larger weight in the regression. This is possible because the squared root of the degrees of freedom was assumed to be the precision factor of this MRA instead of the inverse variance or inverse standard error suggested in MRA literature. (cf. Equation 4.3)
2) WLS 2: The inverse of numbers of tests reported in each primary study is used as weight. This weight is accounting for the large differences in the number of reported estimates per primary study. The weighting factor ensures equal weights per study and thus studies reporting many or few estimates are no longer distorting the analysis. (Geyer-Klingeberg et al. 2018, p. 2175)
3) WLS 3: Finally, a weight of subjective study quality is used. Every study is given one of three weights 1, 2/3 or 1/3 representing the subjective quality estimation of the author of this thesis from good (1) to bad (1/3).

Stanley and Doucouliagos (2017) have shown in simulations that WLS-MRA approaches provide satisfactory estimates, which are sometimes even superior to other approaches.

4.5.4 Outliers

Outliers can distort the linear regression model and should be minimized or omitted. They can be searched based on a composite outlier score obtained via the joint application of multiple outliers detection algorithms (Z-scores (Iglewicz and Hoaglin 1993); interquartile range (IQR); Mahalanobis distance (Cabana et al. 2019); Robust Mahalanobis distance (Gnanadesikan and Kettenring 1972); Minimum Covariance Determinant (Leys et al. 2018); Invariant Coordinate Selection (Archimbaud et al. 2018); OPTICS (Ankerst et al. 1999); Isolation Forest (Liu et al. 2008); and Local Outlier Factor (Breunig et al. 2000)). Outliers are considered as such if they were classified by at least half of the methods used. (Lüdecke 2020)

4.5.5 High leverage points

In contrast to the outliers, where the response *y* is unusual given the predictor *x*, observations with high leverage have an unusual value for x. High leverage observations tend to have a sizable impact on the estimated regression line. Just a couple of high leverage observations may invalidate the entire fit. For this reason, it is important to identify high leverage observations and correct for them. To account for heteroscedasticity and high leverage points at the same time "robust" standard errors can be used. For example the standard errors based on Cribari-Neto and da Silva (2011), with heteroscedasticity consistent and leverage adjusting covariance estimation might be an option. (Zeileis 2004, p. 4f) But in the case of correlation of error terms this approach can't be used anymore, and clustered standard errors must be applied.

4.5.6 Collinearity

Collinearity and multicollinearity are relevant in a multiple linear regression.

Collinearity is detected via a correlation matrix of all moderation variables. Multicollinearity is tested via the variance inflation factor (VIF) from equation (4.7):

Eq. 4.7

$$VIF(\widehat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ the R^2 from a regression of X_j onto all the other predictors. If $R^2_{X_j|X_{-j}}$ is close to 1, then collinearity is present, and so the VIF will be large. In this thesis all VIFs above 5 are assumed to indicate collinearity. There exist two valid approaches to fight collinearity. The first solution is to drop one of the problematic moderation variables from the regression. This usually does not compromise to the regression fit, since the presence of

collinearity implies that the information that this variable provides on the response is redundant in the presence of the other variables. (James et al. 2013, p. 99ff)

Here both approaches are used to gain reliable results.

5 Selection, adaptation and description of the data set

Before conducting the meta analyses the used variables and the data set must be specified and described.

5.1 Selection of relevant data set

From 70 studies and 9,751 coded primary Granger non-causality test results 56 studies and 6,357 results satisfied all the needed restrictions and quality criteria. Those were split into five subsets.

Speculation influencing markets - S2M

The first data set contains all observations where the independent *x* variable is a proxy variable for speculation and the dependent *y* variable is a proxy for the market. This set describes the influence (in a Granger sense) of speculation on the commodity markets and is in the later called "S2M" (speculation to market). The "S2M" subset contains 46 studies with a total of 1,560 observations. The median of test results per study is 112. The study with the most observations includes 374 estimates and the smallest study includes only one test.

Markets influencing speculation - M2S

The reverse data set observing the effect of the markets (*x* variable) on speculation (*y* variable) is in the later called "M2S" (market to speculation). The "M2S" set contains 32 studies with a total of 1,085 observations. The median of test results per study is 117. The study with the most observations includes 374 estimates and the smallest study includes only one test.

Hedging influencing markets – H2M

The data set observing the effect of the hedging (*x* variable) on commodity markets (*y* variable) is called "H2M" (hedging to market). The "H2M" set contains 10 studies with a total of 438 observations. The median of test results per study is 246. The study with the most observations includes 246 estimates and the smallest study includes only one test. The subset contains only few observations and few studies, potential results would be inaccurate. The subset is not used for further analysis.

Markets influencing hedging - M2H

The reverse data set observing the effect of the markets (*x* variable) on hedging (*y* variable) is in the later called “M2H” (market to hedging). The “M2H” set contains 9 studies with a total of 418 observations. The median of test results per study is 246. The study with the most observations includes 246 estimates and the smallest study includes only one test. The same problem from the H2M subset also applies to the M2H subset. The subset is also not used for further analysis.

Remaining observations

The remaining data set of observations not fitting these four data set criteria contains 2,856 observations. Those analyzed for example the relation between two different commodity markets. This subset is also not used in the further analysis.

This thesis only covers speculative influence on markets, hence the data sets M2S, H2M, M2H and the remaining observations are of less importance and omitted in the later analysis.

5.2 Handling of inaccurate or missing data

Some of the primary studies do not contain all the data needed to performed the intended MRA or contain data that has been falsified by rounding errors.

5.2.1 Calculation of missing sample size

In some cases, the sample sizes are not indicated in the primary literature and had to be calculated from the end day minus the start day of the underlying data timeframe and then divided by the periodicity. Sometimes only years or months were indicated, so the start and end day had to be assumed. For start dates the first and for the end dates the last working day of the month or year is selected. It was necessary to calculate the sample size for 72 % of the primary GC tests. The equations used to calculate the missing sample sizes can be seen in appendix A.

5.2.2 Calculation of missing p-values

Some of the p-values were missing but it was possible to calculate them from F, z or Chi^2 results. It was necessary to calculate the p-value for 9 % of the primary GC tests. The equations used to calculate the missing p-values can be seen in appendix A.

5.2.3 Adjustment of inaccurate p-values

As some studies indicate p-values of 0 or 1 two different assumption models were made. Adjusted subset A assumes the p-values of 0 to be very small and p-values of 1 to be very high, therefor all 0 values are assumed to be 1e-180 and all values of 1 to be

0.9999999999999999. Both numbers were chosen since the lowest indicated p-value in all primary literature was 2.97414767539529 e-140 and the highest possible value in R without rounding up to 1 is 0.999982260763392.

A second adjusted subset B assumes the p-values of 0 to be just one decimal point under the value to be rounded up. For example, when the primary study states all p-values with 3 decimal places, the highest p-value resulting in 0 after rounding would be 0.0004, therefor 0.0004 is assumed – ceterus paribus.

The subset B is expected to be more realistic and reliable because it assumes more conservative p-values. Subset A on the other hand can be interpreted as an extreme case. It was necessary to adjust the p-value for 2 % of the primary GC tests.

5.3 Description of meta-regression variables

Several variables, denoted as *Z* in equation 4.6, to measure the impact of different types of heterogeneity are collected. The test characteristics and moderation variables used to explain inconsistent results from the primary studies of the subset S2M are listed in Table 2. All variables are coded manually. The selection of the used moderator variables is driven by the availability from the primary sample, model specification and testing reasons (like (multi-) collinearity) and discussions in the literature.[6]

5.3.1 Commodity groups

91 different commodities were observed in primary literature. Those were grouped into 4 main categories: metals, energy, soft commodities and financials. The fourth category, financials, is omitted because it does not directly represent a raw material, but indices of raw materials and other financial products. It is to be feared that this could falsify results. Figure 2 shows the distribution of commodity groups for the subsamples S2M and M2S.

[6] Over 10 collected variables had to be excluded due to (multi-) collinearity problems.

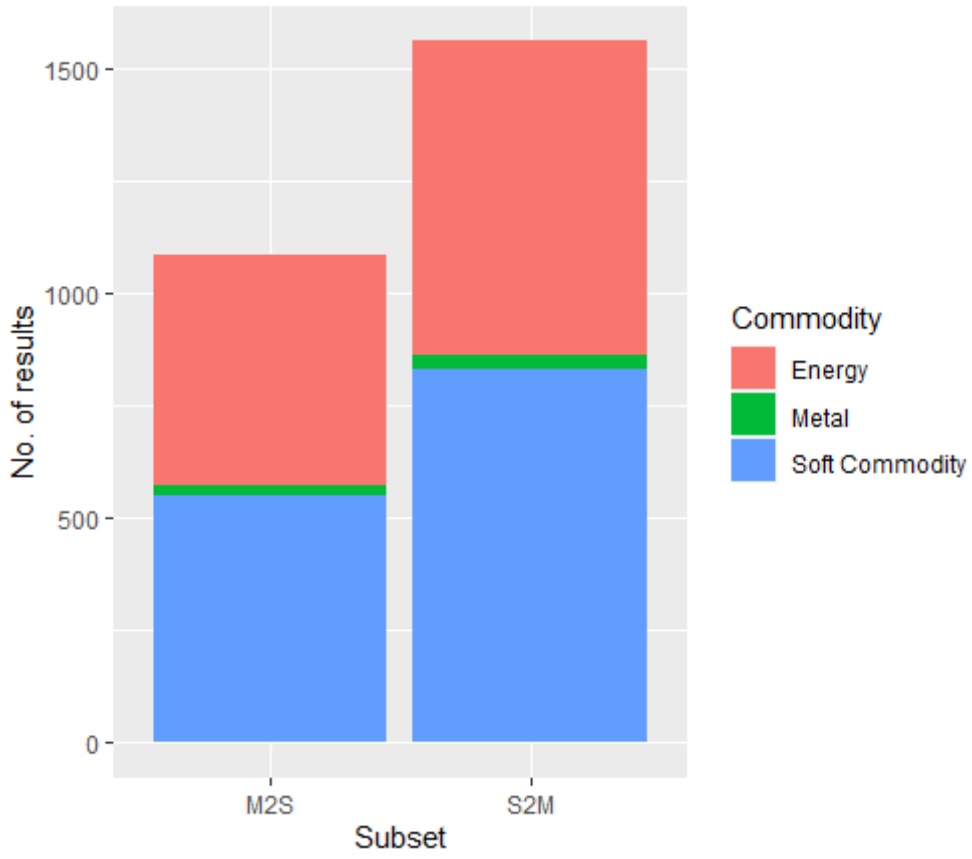

Figure 2: Commodity group distribution of results for subsamples S2M and M2S

Source: Own figure

Metals is the smallest group with only 29 test results from five studies included in the analysis (Ciner 2002, Coleman and Dark 2012, Gilbert and Pfuderer 2014, Gilbert 2009 & Mayer 2012).

5.3.2 Data and test characteristics

In an MRA it must be considered that the underlying studies don't have the same characteristics. All the analyzed GC tests must provide their degrees of freedom as one variable. The degrees of freedom are observed in the primary literature or calculated from the sample size minus lags of *x*, *y* and the lags of potential z-variables. The number of lags for variable *x* of each GC test (Lags M) is observed to investigate potential overfitting of the analyzed model. The different hypothesis test distributions, F-, t- and Chi²-statistic, are each coded as a dummy variable. The underlying time series for *x* variables are in levels, first or more differences. The transformation into difference timeseries is due to the adjustment for stationarity in the underlying time series. The dummy variable "difference" is coded to control for studies that use differenced timeseries compared to those using timeseries in levels. A dummy variable for the logarithmic transformation of the underlying time series, compared to the linear form, are coded. The usage of an autoregressive

distributed lag (ADL) model is the standard case for GC tests. An alternative approach is a GC test using a vector autoregressive regression (VAR) framework and therefore accounting for potential bidirectionality of the relation of speculation and commodity markets. Either the autoregressive distributed lag (ADL) model or the vector autoregressive regression (VAR) / vector error correction (VEC) framework were deployed. The introduced dummy variable 1 if a VAR / VEC framework is used and 0 if a ADL model is deployed. Hypothesis tests normally test for one coefficient to be zero at a time ($H_0: \beta_1 = \cdots = \beta_i = 0$). Testing the sum of coefficients ($H_0: \sum \beta_i = 0$) looks at the persistence of the effects (Gilbert 2010a, p. 41). The testing for coefficient sums is only done in three studies (Gilbert 2010a, Sanders and Irwin 2011a & Sanders et al. 2009). Nevertheless, a dummy variable is added to control for a potential influence. Some of the GC tests used additional control variables (z-variables) to test for influence of factors on the Granger non-causality test. The dummy variable *z-variable* is added to test the influence of those additional moderators. As explained in chapter 2.2 there exist different GC test approaches. Linear and non-linear as well as parametric and non-parametric Granger non-causality tests might produce different results. But GC tests must not always be linear or parametric. The in-quantile GC test is a special method testing for causality in the quantiles or higher means (0.95, 0.8 etc.). Other non-parametric GC tests such as the approaches developed by Hiemstra and Jones (1994), Baek and Brock (1992) and Diks and Panchenko (2006) are coded in a separate dummy variable. Tests using the Baek and Brock (1992) or the Hiemstra and Jones (1994) approach are omitted in the later MRA, because they produce inaccurate and false results. In addition to these linear / non-linear and parametric / non-parametric tests there exist multivariate GC tests. Multivariate GC tests deploy multiple speculation measures in the same test. But due to collinearity no dummy for GC tests in quantiles are investigated. Testing for overfitting via a lag variable answers the question if there is overfitting bias. In most cases the lag structure was selected based on some information criterion. To capture dissimilarities between lag selection based on the AIC and without AIC a dummy (*AICplus*) is introduced that is equal to one for all cases where the AIC approach was used, regardless of the potential simultaneous usage of other information criteria. One of the most common data sources was the CFTC. To check for potential influences by using the same date source this source was coded as a dummy.

5.3.3 Data time characteristics

The data periodicity is observed in daily, weekly, monthly and quarterly steps. The periodicities of the primary studies are coded as unique dummies. The average year is calculated from the start date plus the end date and divided by two. The average year reflects the mean date of the underlying timeseries. The Masters Hypothesis states that in the years after 2000 trading volume and prices were rising rapidly. (Irwin and Sanders

2012, p. 1f) The mean timeframe observed by primary studies is ~7 years. To check if the time after 2000 has an impact, a dummy variable checking for an average year before (one) or after 2007 (zero) is introduced. Due to (multi-) collinearity the dummies for a daily and a quarterly periodicity had to be excluded. Only the *monthly* dummy variable remains in the multiple MRA in chapter 6.4.

5.3.4 Publication characteristics

Some primary studies indicated influence from a third party, like other researchers, scientific institutions, companies or the government. A dummy variable "influence" is constructed for studies where the influence of a third party is stated. Different studies appear in different journals. To evaluate the quality of a scientific work it is common practice taking the journal impact factor into account. In equation 3.1. the calculation of combination of different quality factors is presented. In chapter 6.4.2 the impact of journal ranking on the study precision is investigated in a first analysis. To capture differences between published studies with a high ranking and unpublished studies or studies with a low ranking a dummy (*ranking*) is introduced in the multiple MRA. It is equal to one for all studies in the upper quantile of the ranking factor.

5.3.5 Proxy variable characteristics

In chapter 3.3 different proxy variables for the commodity market and for the speculative activity are introduced. Those proxy variables are linked to the hypothesizes and theories about speculation explained in chapter 2.1. To investigate those hypothesizes three dummy variables representing different proxy variables for commodity market behavior and speculation activity are added to the multiple MRA. The dummy *volatility* is one for all GC tests measuring the influence of speculation on the volatility of commodity prices. It is zero for all GC tests analyzing the influence of speculation on returns, prices and other commodity market proxy variables, like spread or liquidity. For GC tests analyzing the speculation proxy variable OI the dummy variable *OI* is added. A dummy variable *volume* that is equal to one for GC tests analyzing the *y* variable trading volume. If the dummies *OI* and *volume* both are equal to zero, the *y* variable investigated is the position data of speculators together with the remaining other speculation proxy variables.

Table 2: Description and summary of Granger non-causality test characteristics

Variable	Description	Mean	Std. Dev.
Commodity groups			
Metal	= 1 if commodity "metal" is examined, 0 otherwise.	0.02	0.14
Energy	= 1 if commodity "energy" is examined, 0 otherwise.	0.45	0.45
Soft commodity	= 1 if commodity "soft commodity" is examined, 0 otherwise.	0.53	0.50
Data characteristics			
Degrees of freedom	Square root of difference between sample size and included covariates.	851.60	727.01
Lags M	Number of lags of independent variable (x)	3.08	2.73
F-statistic	= 1 if a GC hypothesis test with F-distributed test variable has been performed, 0 otherwise.	0.38	0.49
Chi²-statistic	= 1 if a GC hypothesis test with Chi² distributed test variable has been performed, 0 otherwise.	0.30	0.46
t-statistic	= 1 if a GC hypothesis test with t-distributed test variable has been performed, 0 otherwise.	0.05	0.21
Data in differences	= 1 if underlying x time series is in first or more differences, 0 if underlying x time series is in levels.	0.68	0.47
Log data	= 1 if underlying x time series is logarithmic, 0 if underlying x time series is linear.	0.13	0.33
sum	= 1 if hypothesis test is conducted on sum of betas ($H_0: \sum \beta_i = 0$), 0 if hypothesis test is conducted on single betas ($H_0: \beta_1 = \cdots = \beta_i = 0$)	0.01	0.11
VAR/VEC	= 1 if a VAR/VEC framework is used, 0 if an ADL framework is used.	0.71	0.45
z-variable	= 1 if an additional z-variable is included in the GC test, 0 otherwise.	0.09	0.28

Variable	Description	Mean	Std. Dev.
Linear GC	= 1 if linear GC test is examined, 0 otherwise.	0.85	0.36
Non-parametric GC	= 1 if non-parametric GC test is examined, 0 otherwise.	0.04	0.20
GC in Quantile	= 1 if GC test in quantiles is examined, 0 otherwise.	0.11	0.31
Multivariate GC	= 1 if multivariate GC test is examined, 0 otherwise.	0.003	0.05
AICplus	= 1 if the study author used AIC alone or AIC together with other information criterion to determine the lag length, 0 otherwise.	0.55	0.50
CFTC	= 1 if data of the CFTC were used, 0 otherwise.	0.54	0.50
Data time characteristics			
Daily	= 1 if daily data is examined, 0 otherwise.	0.42	0.49
Weekly	= 1 if weekly data is examined, 0 otherwise.	0.48	0.5
Monthly	= 1 if monthly data is examined, 0 otherwise.	0.10	0.30
Quarterly	= 1 if quarterly data is examined, 0 otherwise.	0.01	0.01
Average year	Average year of data sample ((startyear + endyear)/2)	2005.42	4.12
After 2007	= 1 if average year was after 2007 (((startyear + endyear)/2) – 2007), 0 otherwise.	0.47	0.50
Publication characteristics			
Influenced	=1 if study indicated influence from third party, 0 otherwise.	0.53	0.50
Ranking	= 1 if in upper quantile of calculated “ranking” factor, 0 otherwise.	0.47	0.50
Proxy variable characteristics			
Volatility	= 1 if market proxy variable is volatility, 0 otherwise.	0.258	0.437

Variable	Description	Mean	Std. Dev.
OI	= 1 if speculation proxy variable is OI, 0 otherwise.	0.051	0.219
Volume	= 1 if speculation proxy variable is trading volume, 0 otherwise.	0.067	0.249

Note: "Ranking" is calculated according to equation 3.1. Excluded moderation variables are not listed, unless they are part of an investigated moderator variable group (e.g. periodicity)

6 Empirical results

To determine the impact of speculation several MRA are calculated. Table 2 provides summary statistics for the used MRA variables.

6.1 Graphical investigation of genuine effect and p-hacking

In a first step the complete subset is analyzed. In Figure 3 a scatterplot indicates -probit transformed p-values against the square root of degrees of freedom for subset A and B. The red dotted line indicates the 0.05 significance level. All points above the red dotted line are significant at a 5% level. For a first assessment the linear regression line is displayed in dark blue. The standard deviation is indicated as channel in grey blue.

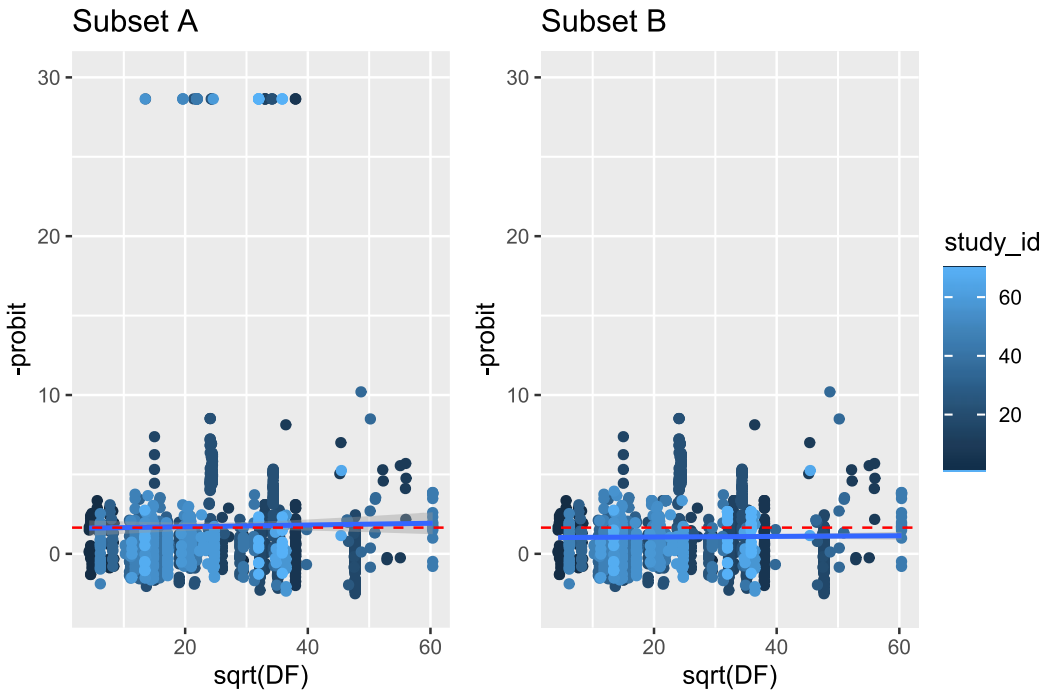

Figure 3: Scatterplot of -probit transformed p-values vs. squared degrees of freedom

Source: Own figure

First statements based on Figure 3 can be made. Subset A has certain outliers in the upper probit-value area introduced by the handling of inaccurate p-values.

To get a better understanding of the data it is helpful to have a look at the density distribution of the probit-values for subset A and B in Figure 4.

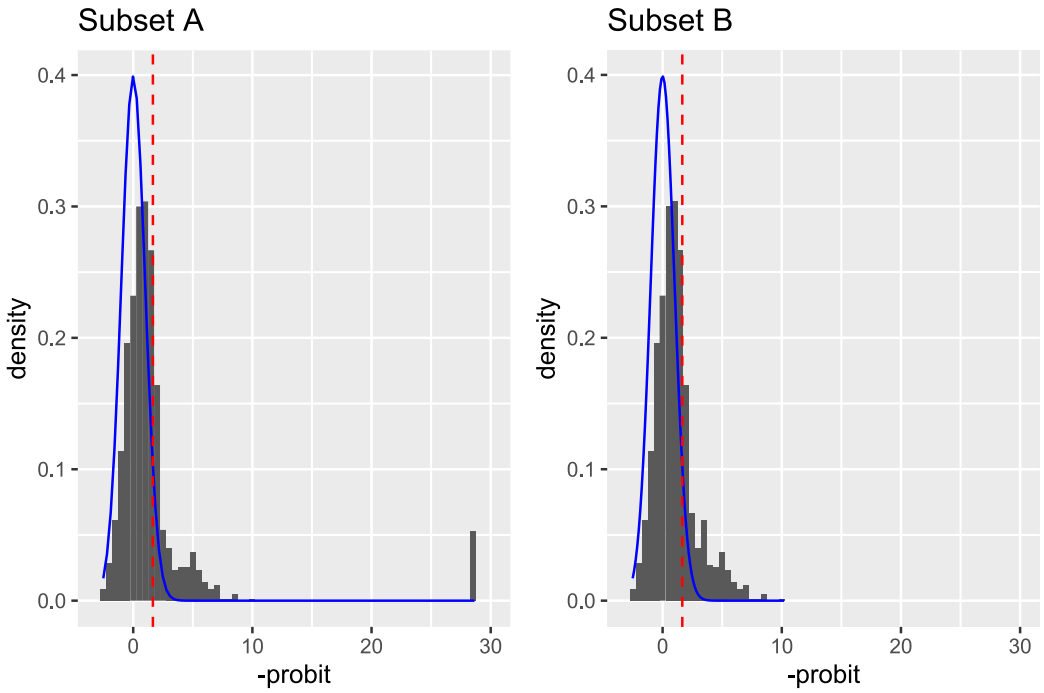

Figure 4: Density distribution of -probit transformed p-values compared to the standard normal distribution

Source: Own figure

The blue curve indicates a standard normal distribution and the red dotted line again indicates the 0.05 significance level.

For subset A the outliers for the adapted p-values of 0 are clearly visible as a bar between the probit-values 25 and 30. For subset B there are no such extreme outliers to the right, but some of the bars right to the significance level are higher. This makes sense as the adapted p-values of 0 are now included in the histogram pillars on the right side. Subset A is still considered to cover the sensitivity for an extreme p-value case. On the left side of the red dotted line both graphs seem to be identical. This also makes sense as only the p-values of 1 and 0 are adapted, and the modifications of p-values of 1 don't differ as extreme as those for p-values of 0. In total, both graphs are skewed to the right with a median of 0.87477 (p-value: 0.191) for both (both have the same) and mean of 1.07173 (p-value: 0.30034) for A and 1.74050 (p-value: 0.30037) for B. Smaller p-values seem overrepresented, but whether the effects are real, come from publication biases or other p-hacking methods cannot be derived from this figure.

To further investigate potential p-hacking the distribution of p-values between 0 and 0.1 is investigated. Figure 5 on the left shows the p-value distribution of all analyzed studies. On the right only studies with a p-value between 0 and 0.1 are investigated. The red dotted line again identifies the 0.05 significance interval. Only subset B is analyzed because the difference between subset A and B is not of interest for the investigation and the results are the same for both subsets.

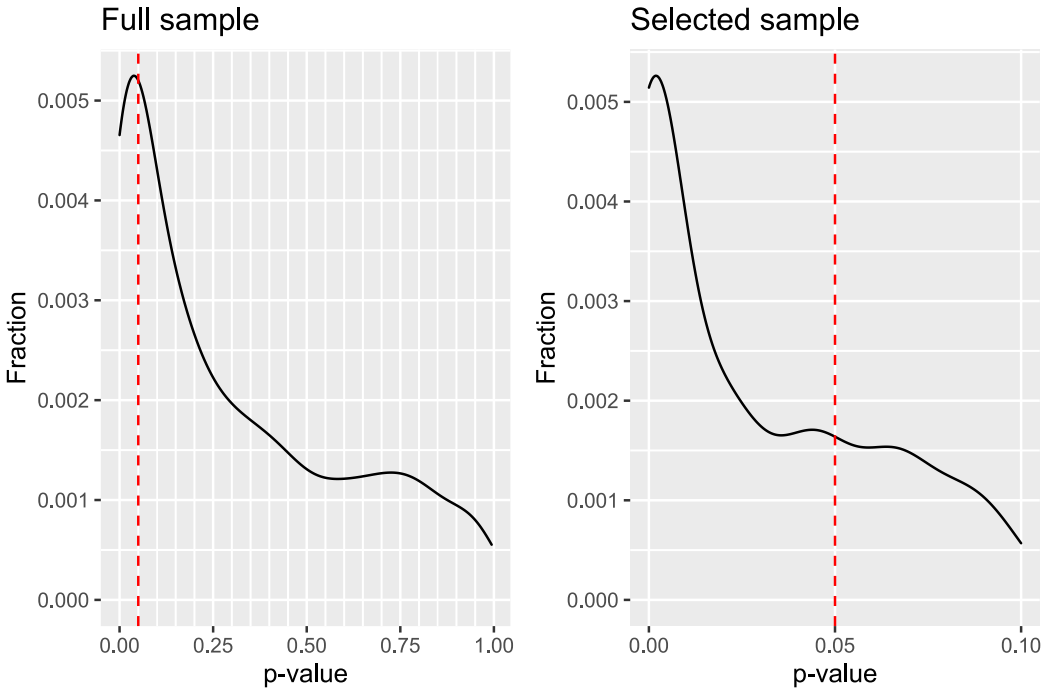

Figure 5: Distribution of p-values for the full studies sample and selected studies sample of p-values between 0 and 0.1

Source: Own figure

When the effect size for a studied phenomenon is zero, it is expected that every p-value is equally likely to be observed. The expected distribution of p-values under the null hypothesis is uniform, such that for the full sample p-value < 0.05 will occur 5% of the time, p-value < 0.04 will occur 4% of the time, and so on. But in Figure 5 it is clearly visible that the bars near the p-value = 0 are higher than the others, indicating there might be a genuine effect. When the true effect is strong, it is more likely to obtain very low p-values (e.g., p-value < 0.001) than moderately low p-values (e.g., p-value < 0.01), and even less likely to obtain nonsignificant p-values (p > 0.05). (Head et al. 2015, p. 3ff)

On both graphs in Figure 5 this right sided downwards slope is clearly visible. A notable drop in p-values above 0.05 is interpreted as evidence for publication bias. While a discontinuity in the distribution of p-values around 0.05 is indicative of publication bias, it does not distinguish between selective publication bias and p-hacking. (Head et al. 2015, p. 3ff)

When there is p-hacking and a truly nonsignificant result is turned into a significant one, then the p-curve's shape will be altered close to the perceived significance threshold (typically p = 0.05). Consequently, a p-hacked plot will have an overabundance of p-values just below 0.05. Both p-hacking and selective publication bias predict a discontinuity in the p-curve around 0.05, but only p-hacking predicts an overrepresentation of p-values just below 0.05. (Head et al. 2015, p. 4ff)

The right graph of Figure 5 shows no apparent pattern in the sample. This suggests that significant results do not stem from p-hacking. But there could be low publication selection as the p-values between 0.05 and 0.1 seem a little bit smaller than those on the left side of 0.05. To gain reliable information about publication selection bias a linear regression model is analyzed in Chapter 6.2

6.2 Testing for publication selection bias

To test explicitly for publication selection bias the results for MRA equation 4.4 are calculated in Table 3. As explained in chapter 4.5 the data set must be tested and adapted to meet certain requirements for linear regression analysis.

Linearity is tested and accepted for all linear regression models. Autocorrelation is determined, hence cluster-robust standard errors are used. A studentized Breusch-Pagan test identifies heteroscedasticity for the most models of subset A and B. Test results are listed in Table 3. Therefore, heteroscedasticity robust cluster standard errors are used and different WLS models are tested for all models. High leverage points and large residuals are identified by calculating Cook's distances and studentized residual. High leverage points with a high influence (large residuals) are excluded individually for each model. In a simple linear regression model no collinearity is observed, therefore no adjustments need to be made. (James et al. 2013, p. 99)

In column 1 the unweighted baseline model is calculated (Base). Column 2 shows the model weighted by the square root of degrees of freedom, to give more precise estimates a higher weight (WLS 1). In column 3 a weighting with the number of tests per study, giving equal weight to each primary study is made (WLS 2). In the last column the weighting is based on a subjective quality assessment (WLS 3). Table 3 shows the obtained coefficients with their t-values, calculated with clustered standard errors on study and author level, in round brackets.

Table 3: Analysis of publication selection bias

	Base	WLS 1	WLS 2	WLS 3
Weight	-	$\sqrt{DF_i}$	$\frac{1}{obs.}$	studyquality
Subset A				
Intercept	**0.997** **(2.3952)***	**1.435** **(2.0194)***	0.234 (1.1387)	**1.107** **(2.5513)***
$\sqrt{DF_i}$	0.000 (0.0143)	-0.015 (-0.8434)	**0.027** **(3.3638)****	-0.003 (-0.1565)
Studentized Breusch-Pagan	0.004	0.128	0.370	0.004
#Studies	46	46	43	46
#Obs	1517	1507	1483	1517
Subset B				
Intercept:	**0.970** **(3.9289)****	**1.159** **(3.0275)****	**0.457** **(2.9330)****	**1.043** **(4.4908)****
$\sqrt{DF_i}$:	-0.001 (-0.1031)	-0.009 (-0.8938)	**0.019** **(2.8686)****	-0.005 (-0.4775)
Studentized Breusch-Pagan	0.1512	0.7568	0.1601	0.1421
#Studies	46	46	42	46
#Obs	1490	1453	1451	1474

Note: This table reports the results of Eq. 4.4. The variable $\sqrt{DF_i}$ refers to the degrees of freedom of the corresponding GC test *i*. Model "Base" in column 1 is unweighted while Model "WLS 1" in column 2 use the square root of the degrees of freedom and "WLS 2" in column 3 the inverse of the number of estimates as weights. "WLS 3" in column 4 uses the subjective study quality evaluation of the author as weights. Model "WLS 2" is the preferred one, because it is the only one considering the different sample sizes per study. The t-statistics of the regression coefficients reported in round brackets are based on clustered standard errors adjusting for within-study and author correlation.

Significance levels are indicated as followed: * p < 0.1, ** p < 0.05, *** p < 0.01

The coefficients of the square-root of degrees of freedom in both subsets are very small and only significant for the "WLS 2" model. Therefore, funnel asymmetry testing (FAT) gives no evidence for the existence or absence of publication selection bias, but for the "WLS 2" model in both subsets. Positive coefficients indicate higher probit-values for

rising degrees of freedom. Higher probit-values retransform to lower p-values and therefore might indicate more significant GC tests results of the primary literature.

Since the coefficients are very small the effect of higher degrees of freedom on the p-values of GC tests is assumed to be small. Hence, distortion due to publication selection bias can be assumed to be small. The FAT has low power, so the results must be taken with caution. The test for precision-effect (PET) shows positive coefficients for subset A and B. For subset A the results are low significant and for model "WLS 2" not significant at all. For subset B the results are significant. (Stanley and Doucouliagos 2016, p. 60ff)

This indicates a potential genuine effect in the underlying GC tests. However, this genuine effect might also have been caused by other biases and distortions. Apart of publication selection bias other biases can influence the results. In order to obtain further information about the primary literature and to account for overfitting bias an augmented MRA model is calculated.

6.3 Testing for publication bias and overfitting bias

The MRA is extended for overfitting bias forming equation 4.5 for which results are shown in Table 4. As before the data set must be tested and adapted to meet certain requirements for linear regression analysis (cf. ch. 4.5). Linearity is tested and accepted for all linear regression models. Autocorrelation is determined, hence cluster-robust standard errors are used. Again, a studentized Breusch-Pagan test is conducted and heteroscedasticity for the most models of subset A and B is identified. Test results of the Breusch-Pagan test are listed in Table 4. Heteroscedasticity robust cluster standard errors are used and different WLS models are tested for all models. The cluster-robust standard error approach and the WLS weighting is used for both subsets to gain comparable results. High leverage points and large residuals are identified by calculating Cook's distances and studentized residual. High leverage points with a high influence (large residuals) are excluded individually for each model. Testing for (multi-)collinearity results that the VIF for all models is smaller 6[7], hence it can be assumed that there is no collinearity.

Again, the basis model (column 1) observes the linear regression without any weighing factor. Model "WLS 1" (column 2) weights the probit-transformed p-values with the square root of degrees of freedom. Model "WLS 2" (column 3) uses the inverse of estimates per study to weight the probit-transformed p-values equally for each primary study. Model "WLS 3" (column 4) uses the subjective quality assessment as weight.

[7]A VIF of 5 or less would be optimal, but to analyze all raw materials the VIF exceeds 5 for a few models. The deviation is only minor and is therefore considered acceptable.

Table 4: Analysis of publication and overfitting bias

Model	Base	WLS 1	WLS 2	WLS 3
Weight	-	$\sqrt{DF_i}$	$\frac{1}{obs.}$	studyquality
Subset A				
Intercept:	**0.976**	1.275	0.213	**1.082**
	(2.0577)*	(1.5206)	(1.0263)	**(2.1562)***
$\sqrt{DF_i}$	0.000	-0.010	**0.020**	-0.003
	(0.0117)	(-0.4683)	**(2.7591)****	(-0.1624)
Lags m	0.012	0.006	0.053	0.017
	(0.4105)	(0.1619)	(1.5530)	(0.5112)
Studentized Breusch-Pagan	0.001	0.0141	0.2721	0.001
#Studies	46	46	43	46
#Obs	1506	1504	1491	1503
Subset B				
Intercept:	**0.914**	**1.184**	**0.356**	**1.009**
	(2.9515)**	**(2.2550)***	**(2.3366)***	**(3.2790)****
$\sqrt{DF_i}$	-0.001	-0.008	**0.016**	-0.003
	(-0.0815)	(-0.7217)	**(2.2435)***	(-0.2304)
Lags m_i	0.021	0.011	**0.066**	0.015
	(0.6127)	(0.2921)	**(2.7958)****	(0.4791)
Studentized Breusch-Pagan	0.0003	0.01428	0.05584	0.009028
#Studies	46	46	40	46
#Obs	1484	1469	1448	1483

Note: This table reports the results of Eq. 4.5. The variable $\sqrt{DF_i}$ refers to the degrees of freedom of the corresponding GC test *i*. Lags m_i refers to the lags of the x variable for test i. Model "Base" in column 1 is unweighted. are estimated by Model 2 – 4 are based on weighted least squares different weights. Model "WLS 1" in column 2 use the square root of the degrees of freedom. "WLS 2" in column 3 applies the inverse of the number of estimates as weights. "WLS 3" in column 4 uses the subjective study quality evaluation of the author as weights. Model "WLS 2" is the preferred one, because it is the only one considering the different sample sizes per study. The t-statistics of the regression coefficients reported in round brackets are based on clustered standard error adjusting for within-study and author correlation.

Significance levels are indicated as followed: * p < 0.1, ** p < 0.05, *** p < 0.01

Adding the lags of speculation (lags m) to the model explains the situation only a little bit better. For publication bias both subsets have positive but small significant coefficients for the "WLS 2" model. A certain, but small publication bias can therefore be assumed. The coefficients for overfitting bias are insignificant except for the "WLS 2" model in subset B. However, the coefficient is positive but small, indicating only low influences on the probit-values. In summary, there are certain indications of publication selection and overfitting biases, however both biases are not expected to strongly influence the GC tests results on speculation effects.

6.4 Analysis of heterogeneity

Publication selection bias and overfitting bias explain the different results and heterogeneity in primary literature only insufficiently. A further analysis must be conducted to find explanation for the excess heterogeneity of primary test results.

6.4.1 Graphical investigation

In a first step different characteristics of the primary studies are analyzed by graphical investigation. In Figure 6 the distribution of p-values of the three commodity groups is plotted.

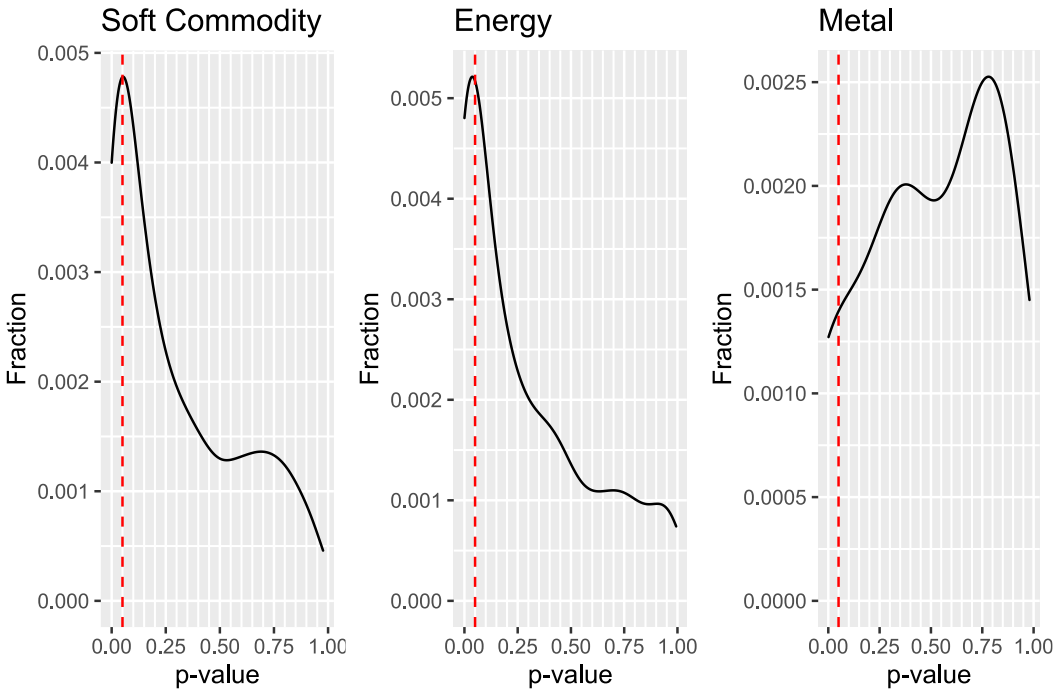

Figure 6: Distribution of p-values of commodity groups

Source: Own figure

In Figure 6 the "L" structure already observed in Figure 5 is visible for soft commodities and energy. This "L" structure can be interpreted as an evidence for genuine effect of

speculation. The graphic for metals is distorted and does not follow the expected course. This could be because only 3 studies have results for metals and hence the sample size is very limited. It is important to pay attention to the different y-axis scales.

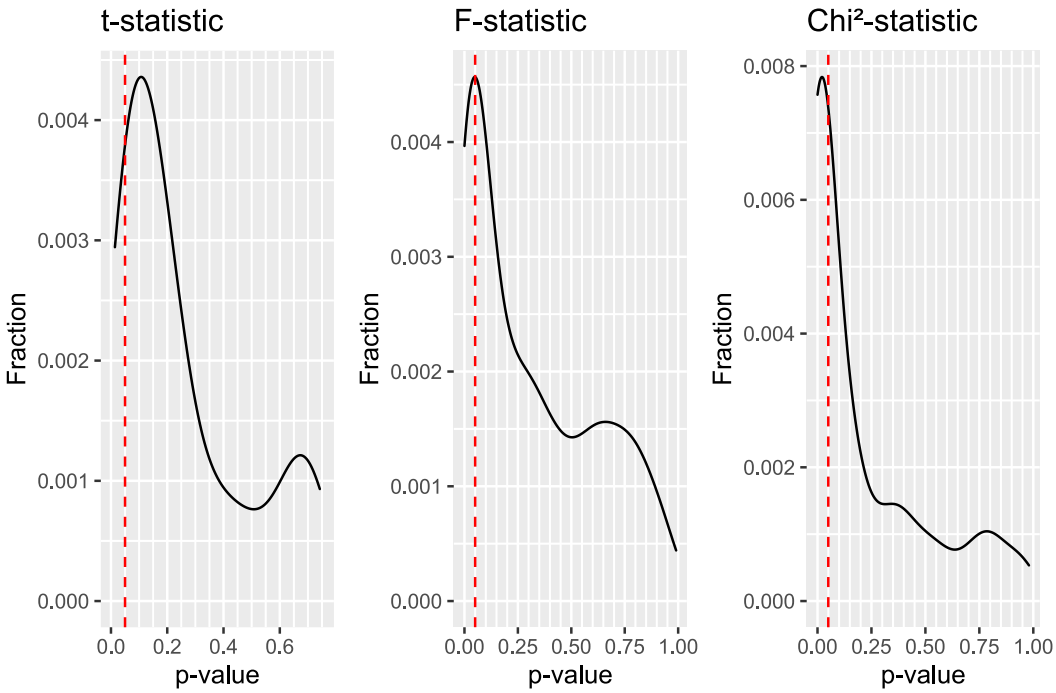

Figure 7: Distribution of p-values of test distribution statistics

Source: Own figure

The p-value distribution of different test distribution statistics can be found in Figure 7.

For all 3 figures the characteristic "L" shape is visible. The t-statistics have their peak on the right side of the significance level, which might indicate potential p-hacking and a lack of significance. The figure of the Chi²-statistic indicates a higher fraction of significant p-values compared to the other two figures.

6.4.2 Journal ranking impact

To check if "higher ranking journals" report more precise estimates Stanley and Doucouliagos (2016) suggest a simple regression of precision against journal quality. Precision is assumed to be the square root of degrees of freedom as stated in chapter 4.2. The ranking of the paper is calculated by equation 3.1 and put into a linear regression model 6.1: (Stanley and Doucouliagos 2016, p. 35)

Eq. 6.1

$$\sqrt{DF_{i,j}} = b_0 + b_1 Ranking_j + u_i$$

Where $DF_{ij}^{0.5}$ refers to the square root of the degrees of freedom of the corresponding GC test *i* of study *j* and $Ranking_i$ is the calculated ranking factor from equation 3.1 for study *j*. The results of equation 6.1 are stated in Table 5. (Stanley and Doucouliagos 2016, p. 35)

Table 5: Literature ranking influence test

Model	DF
Intercept	**19.44** **(41.9) *****
Ranking	**47,520** **(16.63) *****

As coefficient for *Ranking* is highly significant and positive, there is evidence that "higher ranking journals" report more precise estimates. This can be a reason to discard the information contained in the other journals. Nevertheless this study takes a different approach and includes the measure of study quality in the multiple MRA model as another explanatory variable. (Stanley and Doucouliagos 2016, p. 35)

6.4.3 Multiple MRA model

To test more moderation variables and for a more precise analysis a multiple MRA approach is used. The model from equation 4.6 is used. Again, subset A and B are analyzed separately. Linearity is tested and accepted for all linear regression models. Autocorrelation is determined, hence cluster-robust standard errors are used. High leverage points and large residuals are identified by calculating Cook's distances and studentized residual. High leverage points with a high influence (large residuals) are excluded individually for each model. As explained in chapter 4.5.6 collinearity and multicollinearity are a threat for reliable multiple linear regression models. Correlation between potential model parameters is not desired. The model parameters used in Table 6 are the results of a careful selection process testing for correlations and calculating variance inflation factors (VIF) to check for multicollinearity. The selection was made to guarantee the VIF is smaller than 6 for all models. A VIF smaller than 5 would be appreciated, but this would restrict the variable selection drastically. A multiple regression model with a VIF smaller than 6 is therefore accepted as tolerable. Nevertheless, this also means that not all possible moderation variables are included. To identify heteroscedasticity, a studentized Breusch-Pagan test is conducted again. Surprisingly homoscedasticity is identified for most models of both

subsets. Therefor an adjustment for heteroscedasticity is not needed. Test results are listed in Table 6.

Table 6: Analysis of multiple meta regression analysis

	Base	WLS 1	WLS 2	WLS 3	Reduced
	-	$\sqrt{DF_i}$	$\frac{1}{obs.}$	studyquality	$\frac{1}{obs.}$
Subset A					
Intercept	3.053	3.230	1.575	2.757	**1.080**
	(1.6630)	(1.8311)	(1.4044)	(1.3470)	**(5.6365)*****
$\sqrt{DF_i}$	0.043	0.061	-0.006	0.051	
	(1.4402)	(1.8612)	(-0.2007)	(1.5886)	
Lags m:	0.135	0.143	0.234	0.129	
	(1.0731)	(1.0374)	(1.9081)	(1.0294)	
Commodity groups					
Energy			Baseline		
Metal	**-3.037**	**-3.985**	**-2.880**	**-2.602**	**-1.132**
(small sample)	**(-2.4299)***	**(-2.8136)****	**(-2.6832)****	**(-2.1814)***	**(-3.1207)****
Soft Commodities	**-1.787**	**-1.752**	-0.143	**-1.612**	
	(-2.4044)*	**(-2.3297)***	(-0.2142)	**(-2.2909)***	
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	0.086	0.090	0.376	0.152	
	(0.1110)	(0.0892)	(0.5748)	(0.1684)	
t-statistic	1.057	1.269	0.385	1.209	
	(0.7755)	(1.0382)	(0.5010)	(0.7802)	
Differences	0.325	0.247	-0.253	0.360	
	(0.5131)	(0.3954)	(-0.5232)	(0.4909)	
Log x var	-0.828	-0.635	**-1.076**	-0.793	**-0.643**
	(-1.8850)	(-0.9726)	**(-2.3105)***	(-1.7381)	**(-3.0186)****
Sum	0.971	1.370	-1.041	0.841	**-1.093**
	(0.9084)	(1.0107)	(-1.1802)	(0.7410)	**(-4.2513)*****

	Base	WLS 1	WLS 2	WLS 3	Reduced
Var/Vec	-0.669	-0.599	0.014	-0.730	
	(-1.0947)	(-0.8351)	(0.0241)	(-0.9652)	
z-variable	-0.365	-0.566	0.876	-0.047	
	(-0.5044)	(-0.7072)	(1.0963)	(-0.0762)	
Linear GC			Baseline		
non-parametric GC	-1.892	-2.502	-0.955	-2.076	
	(-1.0828)	(-1.4487)	(-0.8504)	(-1.1009)	
Multivariate GC	1.412	1.303	0.034	1.514	
	(1.2240)	(1.2103)	(0.0496)	(1.1696)	
GC in quantile			Excluded due to (multi-) collinearity		
AIC plus	0.686	0.782	-0.095	0.974	
	(0.7822)	(0.7297)	(-0.2171)	(0.8934)	
CFTC	-1.791	-2.026	-0.664	**-2.098**	
	(-1.9413)	(-1.7602)	(-1.0279)	**(-1.9707)***	
Data time characteristics					
Weekly			Excluded due to (multi-) collinearity		
Daily			Excluded due to (multi-) collinearity		
Monthly	0.616	0.874	**0.938**	0.394	
	(1.0729)	(1.1929)	**(1.9701)***	(0.6181)	
quarterly			Excluded due to (multi-) collinearity		
After 2007	0.107	-0.112	0.459	0.159	
	(0.1831)	(-0.1438)	(1.1002)	(0.2475)	
Publication characteristics					
Influenced	-0.821	-1.388	-0.272	-0.751	
	(-1.0075)	(-1.5886)	(-0.5042)	(-0.8468)	
Ranking	-0.120	-0.829	-0.407	-0.017	
	(-0.0959)	(-0.6114)	(-0.6253)	(-0.0122)	
Proxy variable characteristics					
Volatility	-0.610	-0.619	-0.829	-0.657	
	(-1.2251)	(-1.2341)	(-1.5375)	(-1.2074)	

	Base	WLS 1	WLS 2	WLS 3	Reduced
OI	-1.080 (-1.1206)	-1.371 (-1.2627)	-0.474 (-0.5956)	-1.172 (-1.1263)	
Volume	-1.597 (-1.1395)	-2.203 (-1.3940)	0.200 (0.1773)	-1.787 (-1.1667)	
Studentized Breusch-Pagan	1.688e-09	1.229e-09	1.132e-0.08	1.103e-09	0.5077
#Studies	45	44	45	45	44
#Observations	1,501	1,485	1,467	1,496	1,497
Subset B					
Intercept	**2.192** **(2.1663)***	**2.198** **(2.2897)***	1.214 (1.1368)	2.251 (2.2330)*	**1.277** **(3.5135)****
$\sqrt{DF_i}$	0.020 (0.9132)	0.020 (0.9358)	-0.008 (-0.3614)	0.019 (0.8346)	
Lags m:	0.070 (1.2060)	0.103 (1.3480)	**0.154** **(2.7217)****	0.066 (1.1265)	**0.120** **(2.2897)***
Commodity groups					
Energy			Baseline		
Metal (small sample)	**-1.408** **(-3.5981)****	**-1.683** **(-3.1358)****	**-1.430** **(-3.6228)****	**-1.151** **(-2.7125)****	**-1.683** **(-11.2627)****
Soft Commodities	-0.443 (-1.3014)	-0.207 (-0.4305)	0.379 (1.2273)	-0.416 (-1.1914)	
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	-0.318 (-0.6059)	0.044 (0.0715)	0.612 (1.2275)	-0.465 (-0.7748)	
t-statistic	0.026 (0.0442)	0.060 (0.1013	-0.461 (-0.905)	0.066 (0.1092)	
Differences	-0.405 (-1.3252)	-0.419 (-1.1817)	**-0.874** **(-2.0326)***	-0.405 (-1.2002)	**-0.655** **(-2.0516)***

	Base	WLS 1	WLS 2	WLS 3	Reduced
Log x var	**-0.806**	**-0.909**	-0.700	**-0.777**	**-0.755**
	(-3.2002)**	**(-3.6347)*****	(-1.9466)	**(-3.2331)****	**(-3.2378)****
Sum	-0.237	-0.176	-0.533	-0.271	
	(-0.3796)	(-0.2423)	(-0.8840)	(-0.4235)	
Var/Vec	-0.006	-0.129	-0.582	0.164	
	(-0.0134)	(-0.2462)	(-1.3098)	(0.3385)	
Z-variable	-0.063	0.231	0.612	-0.119	
	(-0.1229)	(0.4757)	(0.7417)	(-0.2314)	
Linear GC			Baseline		
Non-parametric GC	0.207	0.570	0.956	0.177	
	(0.3414)	(0.9838)	(1.9433)	(0.2948)	
Multivariate GC	-0.755	-0.744	-1.355	-0.972	**-0.898**
	(-1.1867)	(-1.1237)	(-1.9515)	(-1.4412)	**(-2.6034)****
GC in quantile		Excluded due to (multi-) collinearity			
AIC plus	-0.683	**-0.941**	-0.698	-0.767	
	(-1.6420)	**(-2.0830)***	(-1.5809)	(-1.7326)	
CFTC	-0.721	-0.686	0.088	-0.836	
	(-1.1280)	(-1.0799)	(0.1705)	(-1.2824)	
Data time characteristics					
Weekly		Excluded due to (multi-) collinearity			
Daily		Excluded due to (multi-) collinearity			
Monthly	0.528	0.761	0.730	0.431	
	(0.8171)	(1.0308)	(1.8870)	(0.6310)	
Quarterly		Excluded due to (multi-) collinearity			
After 2007	0.181	0.143	0.346	0.189	
	(0.4226)	(0.2921)	(1.1726)	(0.4314)	
Publication characteristics					
Influenced	-0.433	-0.818	-0.114	-0.478	
	(-0.9747)	(-1.7811)	(-0.3939)	(-1.0369)	

	Base	WLS 1	WLS 2	WLS 3	Reduced
Ranking	0.808 (1.4185)	0.562 (0.9170)	0.308 (0.7197)	0.792 (1.4138)	
Proxy variable characteristics					
Volatility	-0.293 (-0.7695)	-0.454 (-1.0741)	0.002 (0.0065)	-0.307 (-0.7825)	
OI	**-1.317** **(-2.6339)****	**-1.601** **(-3.1594)****	-0.527 (-1.0227)	**-1.377** **(-2.6026)****	
Volume	-0.852 (-0.9486)	-0.890 (-0.9343)	0.733 (0.7040)	-0.833 (-0.8740)	
Studentized Breusch-Pagan	2.2e-16	8.285e-16	2.6e-16	2.2e-16	2.2e-16
#Studies	44	44	44	44	42
#Observations	1,444	1,416	1,405	1,445	1,465

Note: This table reports the results of Eq.4.6 -$probit(p_{ij}) = \beta_1 + \beta_0(DF_{ij}^{0.5}) + \beta_2 lags_{ij} + \sum_{m=1}^{M} \beta_m Z_{mij} + \varpi_{ij}$. The variable $\sqrt{DF_{ij}}$ refers to the degrees of freedom of the corresponding GC test *i* and study *j*. *Lags m* refers to the $lags_{ij}$ of the x variable for test *i* in study *j*. The entries "Metal" till "volume" refer to moderator variables for Z_{mij}. Significance levels are indicated as followed: * p < 0.1, ** p < 0.05, *** p < 0.01

Model "Base" in column 1 is unweighted. Model "WLS 1", "WLS 2" and "WLS 3" are based on weighted least squares with different weights. Model "WLS 1" in column 2 uses the square root of the degrees of freedom. "WLS 2" in column 3 applies the inverse of the number of estimates per study as weights. "WLS 3" in column 4 uses the subjective study quality evaluation of the author as weights. Model "WLS 2" is the preferred one, because it is the only one considering the different sample sizes per study. In column 5 a reduced WLS model only including significant variables is conducted. The t-statistics of the regression coefficients reported in round brackets are based on clustered standard error adjusting for within-study and author correlation.

6.4.4 Reduced model

Stanley and Doucouliagos (2016) suggest a general-to-specific (G-to-S) approach reducing the moderation variables one at a time until only statistically significant variables remain. The G-to-S approach suggests a reduced model where the dummies for soft commodities and energy are excluded and only the dummy for metal is included. But since best practice models are to be created for all commodities, these two dummies are kept in the model. The results of the reduced multiple MRA model is also visible in column 6 of Table 6. (Stanley and Doucouliagos 2016, p. 91ff)

The results of all models including the reduced model are examined in the following chapter.

6.4.5 Results and interpretation

The results for the multiple meta-regression analysis in Table 6 are interpreted for subset A and B. The regression coefficients estimate how the intervention effect in each subgroup differs from the nominated reference subgroup (baseline in this case). The baseline is shown in Table 6 when there are several dummy variables representing different values of the same category (f.e. commodity groups). For dummy variables representing only one moderator variable the baseline is 0 and not indicated. The t-value and p-value of each regression coefficient indicate whether this difference is statistically significant. All coefficients with a p-value above the significance level (0.05) are of less interest. Interpretations of the results only hold for the average of the underlying literature sample of this thesis. In both subsets the coefficients measuring publication selection bias are not significant. This indicates that there are no or no measurable publication selection biases. For subset A no significant coefficients measuring overfitting bias are obtained. For subset B the reduced and the "WLS 2" model both indicate small, significant results. This might be an indication for small overfitting bias, but the effect remains questionable. Both results for publication selection and overfitting biases are mainly in line with the results of the basic MRA and the augmented MRA in chapter 6.2 and 6.3. The coefficients for the metal dummy are negative and significant in all cases. This indicates that on average GC tests analyzing the metal market have higher p-values and therefore are less often significant compared to the energy market. This is in line with the observations in Figure 6. It might indicate that speculation has less effect on the metal market than on other commodities markets. The observed sample of GC tests for the metal market is small, therefore the interpretation is questionable, and results might be distorted. Further research relating the metal markets is needed. For the soft commodities market subset A indicates negative, significant coefficients for the "WLS 1", "WLS 3" and "Base" model. This could indicate that on average GC tests analyzing the soft commodities market have higher p-values and therefore are less often significant compared to the energy market. It might indicate that speculation has less effect on the soft commodities market than on other commodities markets. However, since the results only apply to subset A and are also of very low significance, this interpretation does not seem very realistic. The coefficients for the usage of Chi^2 or t-statistics are all insignificant. This could mean that it is either irrelevant which test statistic is used, or that the F-statistic produces more significant values. It seems reasonable that the selection of a test has no or only a minor effect on the results. The reasons for using first or more differenced time series instead of time series in levels is explained in chapter 2.2. The dummy for differenced time series is low significant and

negative for the “WLS 2” and the reduced model in subset B. This could indicate that on average GC tests on differenced time series have higher p-values and therefore are less often significant compared to the in levels time series. The use of differentiated time series could be interpreted to be more conservative or less accurate. Logarithmic specification of the *x* variable is significant and negative in both subsets. Subset A only produces significant coefficients for the reduced and the “WLS 2” model. For subset B the “Base”, “WLS 1”, “WLS 3” and the reduced model are significant. Using logarithmic data on average results in slightly higher p-values. Just as for the differentiated time series, it could be interpreted that the usage of logarithmic *x* variables is more conservative or less accurate. The use of summed regression coefficients for the GC hypothesis test results significant and negative coefficients for the reduced model in subset A. It is unusual that only the reduced model is significant. Maybe this indicates an interrelation between the moderator variables that has not yet been identified. It could be interpreted as the usage of summed regression coefficients results higher p-values compared to the hypothesis testing of single regression coefficients. This could be interpreted that the hypothesis using the sums is more conservative. But the share of GC tests deploying the summed coefficient approach is rather small and therefor some distortion might exist. The moderator for VAR/VEC is insignificant for all models in both subsets. This means that there is no sign that the usage of the VAR/VEC framework does result in higher or lower p-values compared to the usage of the ADL framework. No significant coefficients for the *z-variable* moderator are obtained, hence there is no indication of different results by GC tests deploying control variables. But results could be distorted because the number of tests with this specification is small (8 %). The primary studies use different GC testing approaches. Due to (multi-)collinearity the dummy variables for the GC test in quantiles and the linear GC test had to be excluded. Dummies for non-parametric and multivariate GC tests are included in the multiple MRA. Only the reduced model for subset B indicates negative and significant coefficients for the multivariate GC testing approach. It is unusual that only the reduced model is significant. Maybe this indicates an interrelation between the moderator variables that has not yet been identified or it is because only a very limited number of tests (0.3 %) deploy multivariate GC tests. The investigation if overfitting bias is influencing the results is already done by the L*ag m* moderator. However, it remains interesting if the usage of the AIC to determine the lag structure influences the results. Only the model “WLS 1” in subset B results low significant and negative coefficients. This could indicate that the usage of AIC to determine the lag structure on average of the underlying studies results higher p-values and hence reject the null hypothesis more often compared to GC tests using no AIC lag length selection. The dummy for CFTC as data source has low significant for “WLS 3” of subset A. From the CFTC data mainly the position data is used in the tests. This could

indicate a more sensitive measurement of speculation by position data than by other data. Nevertheless, only the model weighting by subjective perception of quality returned low significance, which could also be due to random effects. However, further investigations in this area could be useful. Due to (multi-) collinearity the moderation variables for daily, weekly and quarterly had to be excluded. Therefore, only a moderation dummy for a monthly periodicity was included in the MRA. The monthly data is significant only for the "WLS 2" model in subset A. The positive coefficient emphasizes that monthly time series on average of the primary literature sample report lower p-values. This might indicate that speculation on commodity markets can be easier identified in monthly data than in data over different time frames. Because only one model identified low significance and because other periodicity moderators are excluded this result remains questionable. The moderator separating the sample into one part with an average above 2007 and one below 2007 is insignificant for all models of both subsets. This could mean that GC tests with an average year after 2007 report no different p-values compared to those before 2007. An interpretation of these results might be that GC tests with an average after 2007 measure no different influence of speculation on the commodity markets compared to those before 2007. This result would be an argument against Masters hypothesis stating that after 2000 the speculative influence on commodity markets has risen. The dummy variables for the moderating effect of third-party influence on the studies result no significant coefficients. This could mean that influenced studies have no different results compared to not influenced studies. Only 8% of the investigated GC tests take additional control variables into account, the variation of the underlying data might therefore be insufficient. The ranking moderation variable is insignificant, suggesting there is no difference of results in studies published in high ranking journals compared to studies in lower ranking journals or not published at all. That contrasts the results from chapter 6.4.2. Further investigation on the differences between data from studies published in high ranking journals and studies in lower ranking journals or not published studies could be useful.

All three dummy variables representing proxy variables for market behavior and speculation are insignificant for subset A. For subset B the *OI* dummy results negative and significant results for the "Base", the "WLS 1" and "WLS 3" models. This could emphasize higher p-values of GC tests analyzing the influence of *OI* data representing speculation on commodity markets compared to other speculation proxies under investigation. It could be inferred that *OI* as a speculation proxy identifies less speculative influence on commodity markets compared to the other speculation proxies. The *volatility* dummy coefficient is insignificant for both subsets. No differences between GC tests using *volatility* as market behavior proxy and other market proxies like *returns*, prices or other can be recognized. For the *volume* dummy no significant results are obtained either. GC tests based on a

volume variable for *x* does not produce different results compared to the *position of speculators* and other speculation proxy variables. All these results and interpretations are valid for the average of all analyzed primary studies. The interpretations must be taken with caution for the overall situation. Distortions may occur in particular because of a small sample of tests with the investigated property, the omission of high leverage and high influence points or the omission of moderation variables due to (multi-) collinearity. The transfer of the results to reality can only be justified if the basic assumption that the literature sample is representative is accepted.

6.5 Best Practice Model

The previous sub-chapters explain certain biases of the results and heterogeneity of the underlying primary studies. But except for the very simple MRA model in chapter 6.2 they don't state directly if there is an influence of speculation on commodity markets or not. To identify speculation effects a "best practice" model, developed by Stanley and Doucouliagos (2016), is created. The best practice values for the explanatory moderation variables are used and the probit-transformed p-values are predicted. The predicted probit-retransformed p-values are based on the regression coefficients from the "WLS 2" model in Table 7. The best practice model is only calculated for the more conservative subset B. Larger sample sizes produce high degrees of freedom which is desirable for empirical analysis. In this MRA the square root of degrees of freedom is a measurement for publication selection bias. In order to remove identified publication selection bias $\sqrt{DF}$ should be set to 0, but this would distort the results because no study with zero degrees of freedom can be conducted. The mean of degrees of freedom of the underlying studies is used (DF = 850). Following the same logic, it would be wise to remove overfitting bias. However, it makes sense to remain within the scope of the studies examined, because selecting a lag length of zero might result in underfitting bias. Therefore, the mean of 3 lags is used for the best practice model. This is not perfect, because the regression analysis indicated a small overfitting bias and this bias might be introduced into the best practice model hereby, but no better approach to define the lag moderator is available. A F-statistic is used for the statistical test distribution. From the primary literature it is known that speculation variables in first or more differences are often more reliable regarding stationarity. (Malliaris and Urrutia 1998, p. 62ff). Therefore, differenced time series structure is used. Most tests use linear variable structure instead of logarithmic transformed ones. Linear variable structure is therefore used in the best practice model. To prevent omitted variable bias, it seems reasonable to include control variables in GC tests. But primary literature reports only limited usage of additional variables (Shanker 2017; Etienne et al. 2017). Hence the mean of the z-variable moderator is inserted in the best practice

equation (*z-variable* = 0.08077). Standard, linear GC tests are the most used GC testing method. When standard GC tests are used pre-tests for the absence of cointegration between the variables are required. Alternatively, a robust GC test, like the Dolado and Lütkepohl 1996 Granger causality tests, that does not require a pre-test for cointegration, could be applied as described before. Both, standard linear and robust GC tests, are included in the moderator *linear GC*. Linear GC testing is assumed to gain reliable and comparable results. The usage of AIC should be omitted due to the risk of overfitting bias (*AICplus* = 0). For the dummy *CFTC* the mean is assumed, because it should be irrelevant if CFTC data is used or not (*CFTC* = 0.54). *Weekly* should be used for periodicity, because it is used in most of the primary studies, But the dummies *weekly*, *quarterly* and *daily* had to be excluded due to collinearity. Therefore, the dummy *monthly* is set to 0. This is not perfect but the best approach available. Ideally studies should not be influenced by corporations or governments to guarantee scientific objectivity. The moderator for influence however does not separate between different third parties. The mean of the *influenced* moderator is used. (*influenced* = 0.53). The journal ranking impact test in chapter 6.4.2 indicates a certain influence of journal ranking, but the multiple MRA in Table 6 results no influence of high ranking studies. To test if high ranking studies have different p-values the best practice model is conducted for both dummy cases. All three commodities are of interest therefore different cases are calculated for each of them. The hypothesizes presented in chapter 2.1 are based in different proxy variables. No decision can be made if one proxy variable approach for speculation and market behavior proxies is better than the other. For this reason and to be able to assess the different results of the different approaches different cases for all three dummy variables (*volatility*, *volume*, *OI*) are calculated. To also test for the Masters hypothesis cases for an average year before 2007 and after 2007 are also calculated. Overall, the calculation of the best practice approach is therefore a calculation of several different approaches. However, this is assumed to be justified as a clear postulation of one best price model is not possible at the current state of knowledge and research.

Table 7: Best practice model for S2M, subset B, model WLS 2 influencing commodity market volatility

			High ranking			Low ranking		
			Soft commodity	Energy	Metal	Soft commodity	Energy	Metal
Volatility	Volume	Avg year before 2007	**0.073***	0.141	0.638	0.126	0.222	0.746
		Avg year after 2007	**0.036****	**0.078***	0.503	**0.068***	0.133	0.624
	OI	Avg year before 2007	0.423	0.573	0.947	0.546	0.689	0.973
		Avg year after 2007	0.295	0.436	0.898	0.408	0.558	0.943
	Position of speculators / other	Avg year before 2007	0.235	0.366	0.862	0.340	0.486	0.919
		Avg year after 2007	0.143	0.246	0.771	0.224	0.352	0.853
Return / price / other	Volume	Avg year before 2007	**0.073***	0.142	0.639	0.126	0.222	0.747
		Avg year after 2007	**0.036****	**0.078***	0.504	**0.068***	0.133	0.625
	OI	Avg year before 2007	0.424	0.574	0.947	0.547	0.690	0.973
		Avg year after 2007	0.295	0.437	0.898	0.409	0.559	0.943
	Position of speculators / other	Avg year before 2007	0.236	0.367	0.862	0.341	0.487	0.919
		Avg year after 2007	0.143	0.246	0.771	0.225	0.353	0.854

Notes: This table shows calculated hypothetical p-values for best practice models. Different cases are calculated for the three commodity groups: *soft commodities*, *energy* and *metal*. *Avg year before 2007* refers to a average sample period year before 2007. *Avg year after 2007* refers to an average sample period year after 2007. *High ranking* refers to studies in the upper quantile of the calculated ranking factor in chapter 6.4.2. *Low ranking*

identifies all studies not in the upper quantile. *Volume, OI* and *position of speculators / other* are related to studies applying them respectively as speculation proxy variables. *Volatility* and *return / price / other* are related to studies applying them respectively as market behavior proxy variables.

The obtained predictions in Table 7 are the mean p-values for different hypothetical studies deploying different versions of best practice specifications. In line with the results of the multiple MRA the results in Table 7 show that GC tests in the metal market cannot reject GC in any model setup. For studies investigating the soft commodities and energy commodities market this finding holds in large parts. Only 8 model setups out of the 72 deliver significant GC results and hence predict that speculation significantly Granger-causes commodity market volatility, returns, etc. The best practice approaches for soft commodities in high ranking studies with the *volume* speculation proxy return significant p-values at a significance level of 5 % for an average sample year after 2007 and of 10% for an average sample year before 2007. This could indicate that the trading volume has a certain influence on the soft commodity market. The results could also support the Masters hypothesis under certain circumstances for the soft commodity market. For the energy market the best practice approach in high ranking studies with the *volume* speculation proxy and an average sample year after 2007 returns significant p-values with a significance level of 10 %. Again, significant results are only obtained for the *volume* proxy variable. One could interpret the situation as meaning that a change in trading volume has an impact on volatility, price etc. of soft and energy commodities. Following *Friedman's speculative stabilizing theory* and the *risk-transfer hypothesis* higher trading volume should mean more liquidity, more executed trades, lower OI and therefore more stable prices and lower volatility. But the obtained results fit better with the *noise-trader hypothesis* and the *bull-and-bear hypothesis*. However, significant results would then be expected for *OI* and *position of speculators*. This raises the question if the results for the *volume* proxy are due to speculation or due to some other effects. The obtained significant results might occur because of misinterpretation of the underlying interrelations of the commodity market variables. It should be noted, that the p-values for tests after 2007 are always lower than those before 2007. This would support Masters hypothesis that the influence of speculation activity has increased in the years after 2000. Overall 1,409 GC tests from 43 primary studies find only limited and questionable evidence for speculation effects on commodity markets. Those results might imply that speculation in general does not influence commodity markets or that the underlying study and test design are of low power. Additionally, the availability of meaningful and reliable data to conduct GC tests is challenging. The availability of meaningful and reliable primary studies is even more challenging. The broad diversity of different applied measures for speculation and market behavior as well as the different testing approaches make it difficult to aggregate and

analyze them. Especially the absence of a perfect measurement method to identify speculation makes it hard to gain reliable results.

7 Further research and outlook

The present MRA approach only represents one possibility to analyze Granger non-causality tests. Many different approaches are possible. For example, the working paper of Wimmer et al. 2020 uses a panel Granger approach and a classical MRA model. Like the methods used in this thesis they use the weighting factors square-root of degree of freedom and inverse of the number of estimates reported in each primary study. However, they supplement them with a third weighting factor. This third factor is the inverse of the approximated GC test variance, as suggested by Dumitrescu and Hurlin (2012). The calculations can be seen in appendix C and D and can be compared to the main results.

As explained in chapter 4.5, correlation, heteroscedasticity and high leverage points are of relevance for the validity and reliability of the analysis. The used MRA analysis adequately corrects for correlation and heteroscedasticity, but a correction of the high leverage points without deleting them might be useful. An experimental approach using two way clustered standard errors calculated with heteroscedasticity consistent and high leverage correcting covariance estimation following MacKinnon and White 1985 exists. This approach is not included in the main results, because it is considered experimental by Zeileis (2004). Nonetheless, the calculations were conducted for the full model and included in appendix D.

Further investigation of trading effects, for example a comparison between the effects of speculation and those of hedging activity on the markets and potential influence of the markets on those activities, should be done, especially since the data basis has already been created by the coding work of this thesis.

In a further investigation of meta regression analysis of commodity markets and speculation it would be reasonable to model interaction hypothesizes with the additional predictors to further gain insights. The analysis of moderation variables excluded in this MRA approach could provide more answers.

During this thesis several, different permutations of the multiple MRA were calculated. It is noticeable that the results have sometimes changed considerably even with small changes in the model setup. In the further course of research, it could be useful to perform permutation tests of the model to further validate the robustness of the statistical models. The permutation test is a special form of resampling methods. It assesses if the model

captures a true pattern underlying the primary data. That way potential overfitting of the MRA model can be investigated. (Good 2000, p. 128; Harrer et al. 2019, ch. 83.)

8 Conclusion

This thesis applies meta-regression analysis to aggregate and systematically analyze 46 empirical studies on the impact of speculation on commodity markets. A total of 1,560 Granger non-causality test results is analyzed with the Meta-GC model proposed by Bruns and Stern (2015).

In a first step the underlying literature sample is tested for publication selection and over-/underfitting by lag selection. A basic MRA and an augmented MRA model find only very small evidence for low publication selection bias and overfitting bias.

In a second step a multiple meta-regression analysis with moderator variables capturing various aspects of primary study characteristics and test design is conducted. The results suggest that GC from speculation is less present in metal markets compared to energy and soft commodity markets. GC tests using differenced or logarithmic transformed time series report less often significant results. OI as speculation proxy variable is less likely to present significant p-values compared to the other speculation proxy variables.

A best practice model is derived from the multiple MRA model. The hypothesis of Granger non-causality between speculation and commodity markets cannot be rejected significantly for most calculated cases at standard significance levels. For soft and energy commodities some of the results suggest a rejection of the hypothesis that trading volume is not Granger causing market behavior. Together with the lower p-values for GC tests with an average year after 2007 this supports the Masters hypothesis of higher financialization of commodity markets in the recent years. Studies with a high journal ranking, much citations etc. are more likely to result significant results compared to studies with a lower ranking or no ranking.

Based on the MRA models and the best practice model, there is only little evidence for speculative influence on the markets for soft commodities and energy. In summary little to no genuine overall speculation effect is detectable for soft commodity, energy and metal markets on the average of the investigated primary literature.

9 Publication bibliography

Abdullahi, Abba Saada; Kouhy, Reza; Muhammad, Zahid (2014): Trading volume and return relationship in the crude oil futures markets. In *Studies in Economics & Finance* 31 (4), pp. 426–438. DOI: 10.1108/SEF-08-2012-0092.

Akaike, H. (1974): A new look at the statistical model identification. In *IEEE Trans. Automat. Contr.* 19 (6), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

Algieri, Bernardina (2016): Conditional price volatility, speculation, and excessive speculation in commodity markets: sheep or shepherd behaviour? In *International Review of Applied Economics* 30 (2), pp. 210–237. DOI: 10.1080/02692171.2015.1102204.

Alquist, Ron; Gervais, Olivier (2013): The Role of Financial Speculation in Driving the Price of Crude Oil. In *EJ* 34 (3). DOI: 10.5547/01956574.34.3.3.

Amann, Stefan; Lehecka, Georg V.; Schmid, Erwin (2013): Does speculation drive agricultural commodity spot prices? Treibt Spekulation agrarische Kassapreise? In *Jahrbuch der Österreichischen Gesellschaft für Agrarökonomie* 2013 (22), pp. 131–140. Available online at https://oega.boku.ac.at/fileadmin/user_upload/Tagung/2012/Band_22_1/12_Amann_et_al_OEGA_Jahrbuch_2012.pdf.

Amato, Massimo; Cavalli, Nicolò; Cifarelli, Giulio; Cristiano, Carlo; Fantacci, Luca; Foresti, Tiziana et al. (2012): Speculation and Regulation in Commodity Markets: The Keynesian Approach in Theory and Practice. Edited by Sapienza University of Rome. Sapienza University of Rome (Rapporto Tecnico, 21).

Ang, James B. (2008): A survey of recent developments in the literature of finance and growth. In *J Economic Surveys* 22 (3), pp. 536–576. DOI: 10.1111/j.1467-6419.2007.00542.x.

Ankerst, Mihael; Breunig, Markus M.; Kriegel, Hans-Peter; Sander, Jörg (1999): OPTICS. In *SIGMOD Rec.* 28 (2), pp. 49–60. DOI: 10.1145/304181.304187.

Antonakakis, Nikolaos; Chang, Tsangyao; Cunado, Juncal; Gupta, Rangan (2018): The relationship between commodity markets and commodity mutual funds: A wavelet-based analysis. In *Finance Research Letters* 24, pp. 1–9. DOI: 10.1016/j.frl.2017.03.005.

Arai, Mahmood (2015): Cluster-robust standard errors using R. Department of Economics, Stockholm University.

Archimbaud, Aurore; Nordhausen, Klaus; Ruiz-Gazen, Anne (2018): ICS for multivariate outlier detection with application to quality control. In *Computational Statistics & Data Analysis* 128, pp. 184–199. DOI: 10.1016/j.csda.2018.06.011.

Aulerich, Nicole M.; Irwin, Scott H.; Garcia, Philip (2010): The Price Impact of Index Funds in Commodity Futures Markets: Evidence from the CFTC's Daily Large Trader Reporting System. conference paper.

Aulerich, Nicole M.; Irwin, Scott H.; Garcia, Philip (2014): Bubbles, Food Prices, and Speculation Evidence from the CFTC's Daily Large Trader Data Files. Evidence from the CFTC's Daily Large Trader Data Files. In *The Economics of Food Price Volatility*, 211-253. Available online at http://www.nber.org/chapters/c12814.

Babalos, Vassilios; Balcilar, Mehmet (2017): Does institutional trading drive commodities prices away from their fundamentals: Evidence from a nonparametric causality-in-quantiles test. In *Finance Research Letters* 21, pp. 126–131. DOI: 10.1016/j.frl.2016.11.017.

Baek, E.; Brock, W. (1992): A general test for Granger causality: bivariate model. In *TechnicalReport*.

Baldi, Lucia; Peri, MassimoM; Vandone, Daniela (2011): Spot and Futures Prices of Agricultural Commodities: Fundamentals and Speculation. Proceedings in Food System Dynamics, Proceedings in System Dynamics and Innovation in Food Networks. conference paper, pp. 110–125. DOI: 10.18461/pfsd.2011.1110.

Bell, David; Kay, Jim; Malley, Jim (1996): A non-parametric approach to non-linear causality testing. In *Economics Letters* 51 (1), pp. 7–18. DOI: 10.1016/0165-1765(95)00791-1.

Bessembinder, Hendrik; Seguin, Paul J. (1992): Futures-Trading Activity and Stock Price Volatility. In *The Journal of Finance* 47 (5), pp. 2015–2034. DOI: 10.1111/j.1540-6261.1992.tb04695.x.

Bohl, Martin T.; Siklos, Pierre L.; Wellenreuther, Claudia (2018): Speculative Activity and Returns Volatility of Chinese Major Agricultural Commodity Futures. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.3105109.

Borin, Alessandro; Di Nino, Virginia (2012): The Role of Financial Investments in Agricultural Commodity Derivatives Markets. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.2030780.

Bos, Jaap W.B.; van der Molen, Maarten (2012): A Bitter Brew? Futures Speculation and Commodity Prices. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.2209706.

Breunig, Markus M.; Kriegel, Hans-Peter; Ng, Raymond T.; Sander, Jörg (2000): LOF. Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00, pp. 93–104. DOI: 10.1145/342009.335388.

Brümmer, Bernhard; Korn, Olaf; Schlüßler, Kristina; Jaghdani, Tinoush Jamali; Saucedo, Alberto (2013): Volatility in the after crisis period – A literature review of recent empirical research. Working Paper No. 1. ULYSSES "Understanding and coping with food markets voLatility towards more Stable World and EU food SystEmS". Available online at https://www.academia.edu/download/42649919/ULYSSES_Working_Paper_1_Volatility_in_the_after_crisis_period_-_A_literature_review_of_recent_empirical_research5.pdf.

Brunetti, Celso; Buyuksahin, Bahattin (2009): Is Speculation Destabilizing? working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.1393524.

Brunetti, Celso; Buyuksahin, Bahattin; Harris, Jeffrey H. (2011): Speculators, Prices and Market Volatility. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.1736737.

Brunetti, Celso; Buyuksahin, Bahattin; Harris, Jeffrey H. (2013): Herding and Speculation in the Crude Oil Market. In *EJ* 34 (3). DOI: 10.5547/01956574.34.3.5.

Bruns, Stephan B.; Gross, Christian; Stern, David I. (2014): Is There Really Granger Causality Between Energy Use and Output? In *EJ* 35 (4). DOI: 10.5547/01956574.35.4.5.

Bruns, Stephan B.; Stern, David I. (2015): Meta-Granger Causality Testing. In *SSRN Journal*. DOI: 10.2139/ssrn.2619478.

Bruns, Stephan B.; Stern, David I. (2019): Lag length selection and p-hacking in Granger causality testing: prevalence and performance of meta-regression models. In *Empir Econ* 56 (3), pp. 797–830. DOI: 10.1007/s00181-018-1446-3.

Bu, Hui (2011): Price Dynamics and Speculators in Crude Oil Futures Market. In *Systems Engineering Procedia* 2, pp. 114–121. DOI: 10.1016/j.sepro.2011.10.014.

Buyuksahin, Bahattin; Harris, Jeffrey H. (2011): Do Speculators Drive Crude Oil Futures Prices? In *EJ* 32 (2). DOI: 10.5547/issn0195-6574-ej-vol32-no2-7.

Cabana, Elisa; Lillo, Rosa E.; Laniado, Henry (2019): Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. In *Stat Papers* 141 (2), p. 817. DOI: 10.1007/s00362-019-01148-1.

Cameron, A. Colin; Gelbach, Jonah B.; Miller, Douglas L. (2011): Robust Inference With Multiway Clustering. In *Journal of Business & Economic Statistics* 29 (2), pp. 238–249. DOI: 10.1198/jbes.2010.07136.

Capelle-Blancard, Gunther; Coulibaly, Dramane (2012): Index Trading and Agricultural Commodity Prices: A Panel Granger Causality Analysis. In *SSRN Journal*. DOI: 10.2139/ssrn.1980058.

Card, David; Krueger, Alan B. (1995): Time-Series Minimum-Wage Studies: A Meta-analysis. In *The American Economic Review* (85), pp. 238–243.

Chakraborty, Ranajit; Das, Rahuldeb (2013): Dynamic Relationship Between Futures Trading and Spot Price Volatility: Evidence from Indian Commodity Market. In *IUP Journal of Applied Finance* (19).

Chuang, Chia-Chang; Kuan, Chung-Ming; Lin, Hsin-Yi (2009): Causality in quantiles and dynamic stock return–volume relations. In *Journal of Banking & Finance* 33 (7), pp. 1351–1360. DOI: 10.1016/j.jbankfin.2009.02.013.

Ciner, Cetin (2002): Information content of volume: An investigation of Tokyo commodity futures markets. In *Pacific-Basin Finance Journal* 10 (2), pp. 201–215. DOI: 10.1016/s0927-538x(01)00037-3.

Coleman, Les; Dark, Jonathan (2012): Economic Significance of Non-Hedger Investment in Commodity Markets. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.2021919.

Cribari-Neto, Francisco; da Silva, Wilton Bernardino (2011): A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. In *AStA Adv Stat Anal* 95 (2), pp. 129–146. DOI: 10.1007/s10182-010-0141-2.

Deb, Partha; Trivedi, Pravin K.; Varangis, Panayotis (1996): The excess co-movement of commodity prices reconsidered. In *J. Appl. Econ.* 11 (3), pp. 275–291. DOI: 10.1002/(SICI)1099-1255(199605)11:3<275::AID-JAE392>3.0.CO;2-3.

Diks, Cees; Panchenko, Valentyn (2005): A Note on the Hiemstra-Jones Test for Granger Non-causality. In *Studies in Nonlinear Dynamics & Econometrics* 9 (2). DOI: 10.2202/1558-3708.1234.

Diks, Cees; Panchenko, Valentyn (2006): A new statistic and practical guidelines for nonparametric Granger causality testing. In *Journal of Economic Dynamics and Control* 30 (9-10), pp. 1647–1669. DOI: 10.1016/j.jedc.2005.08.008.

Ding, Haoyuan; Kim, Hyung-Gun; Park, Sung Y. (2014): Do net positions in the futures market cause spot prices of crude oil? In *Economic Modelling* 41, pp. 177–190. DOI: 10.1016/j.econmod.2014.05.008.

Dolado, Juan J.; Lütkepohl, Helmut (1996): Making wald tests work for cointegrated VAR systems. In *Econometric Reviews* 15 (4), pp. 369–386. DOI: 10.1080/07474939608800362.

Dumitrescu, Elena-Ivona; Hurlin, Christophe (2012): Testing for Granger non-causality in heterogeneous panels. In *Economic Modelling* 29 (4), pp. 1450–1460. DOI: 10.1016/j.econmod.2012.02.014.

Ederer, Stefan; Heumesser, Christine; Staritz, Cornelia (2013): The role of fundamentals and financialisation in recent commodity price developments: An empirical analysis for wheat, coffee, cotton, and oil. working paper. Edited by econstor. Austrian Foundation for Development Research (ÖFSE), Vienna (42).

Etienne, Xiaoli L.; Irwin, Scott H.; Garcia, Philip (2017): New Evidence that Index Traders Did Not Drive Bubbles in Grain Futures Markets. In *Journal of Agricultural and Resource Economics* 2017 (42), pp. 45–67.

Fagan, Stephen; Gencay, Ramazan (2008): Liquidity-Induced Dynamics in Futures Markets. working paper, pp. 1–38. Available online at http://mpra.ub.uni-muenchen.de/6677/.

Figlewski, Stephen (1981): Futures Trading and Volatility in the GNMA Market. In *The Journal of Finance* 36 (2), pp. 445–456. DOI: 10.1111/j.1540-6261.1981.tb00461.x.

Friedman, Milton (1953): Essays in positive economics. Chicago: Univ. of Chicago Press.

Fujihara, Roger A.; Mougou, Mbodja (1997): An examination of linear and nonlinear causal relationships between price variability and volume in petroleum futures markets. In *J. Fut. Mark.* 17 (4), pp. 385–416. DOI: 10.1002/(sici)1096-9934(199706)17:4%3C385::aid-fut2%3E3.0.co;2-d.

Fürst, Wilhelm (1896): Börsengesetz. Vom 22. Juni 1896. In *Jahrbücher für Nationalökonomie und Statistik* 67 (1). DOI: 10.1515/jbnst-1896-0126.

Geyer-Klingeberg, Jerome; Hang, Markus; Rathgeber, Andreas W. (2019): What drives financial hedging? A meta-regression analysis of corporate hedging determinants. In *International Review of Financial Analysis* 61, pp. 203–221. DOI: 10.1016/j.irfa.2018.11.006.

Geyer-Klingeberg, Jerome; Hang, Markus; Walter, Matthias; Rathgeber, Andreas (2018): Do stock markets react to soccer games? A meta-regression analysis. In *Applied Economics* 50 (19), pp. 2171–2189. DOI: 10.1080/00036846.2017.1392002.

Ghalayini, Latife (2011): The Interdependence of Oil Spot and Futures Markets. In *European Journal of Economics, Finance and Administrative Sciences* (32), pp. 1–13.

Ghalayini, Latife (2017): Modeling and forecasting spot oil price. In *Eurasian Bus Rev* 7 (3), pp. 355–373. DOI: 10.1007/s40821-016-0058-0.

Gilbert, Christopher L. (2009): Speculative Influences on Commodity Futures Prices 2006-08.

Gilbert, Christopher L. (2010a): Commodity speculation and commodity investment. In *Commodity market review 2009-2010*.

Gilbert, Christopher L. (2010b): How to Understand High Food Prices. In *J Agric Econ* 61 (2), pp. 398–425. DOI: 10.1111/j.1477-9552.2010.00248.x.

Gilbert, Christopher L.; Pfuderer, Simone (2012): Index Funds Do Impact Agricultural Prices. workshop paper. Edited by Department of Economics, University of Trento.

Gilbert, Christopher L.; Pfuderer, Simone (2014): The Role of Index Trading in Price Formation in the Grains and Oilseeds Markets. In *J Agric Econ* 65 (2), pp. 303–322. DOI: 10.1111/1477-9552.12068.

Girardi, Daniele (2015): Financialization of food. Modelling the time-varying relation between agricultural prices and stock market dynamics. In *International Review of Applied Economics* 29 (4), pp. 482–505. DOI: 10.1080/02692171.2015.1016406.

Gnanadesikan, R.; Kettenring, J. R. (1972): Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. In *Biometrics* 28 (1), pp. 81–124. DOI: 10.2307/2528963.

Gonzalo, Jesus; Pitarakis, Jean-Yves (2002): Lag length estimation in large dimensional systems. In *J Time Series Analysis* 23 (4), pp. 401–423. DOI: 10.1111/1467-9892.00270.

Good, Phillip (2000): Permutation Tests. New York, NY: Springer New York.

Granger, C. W. J. (1969): Investigating Causal Relations by Econometric Models and Cross-spectral Methods. In *Econometrica* 37 (3), p. 424. DOI: 10.2307/1912791.

Greene, William H. (2003): Econometric analysis. 5th ed., International ed. Upper Saddle River, N.J., [Great Britain]: Prentice Hall.

Gupta, C. P.; Sehgal, Sanjay; Wadhwa, Sahaj (2018): Agricultural Commodity Trading: Is it Destabilizing Spot Markets? In *Vikalpa* 43 (1), pp. 47–57. DOI: 10.1177/0256090917750263.

Gusenbauer, Michael (2019): Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. In *Scientometrics* 118 (1), pp. 177–214. DOI: 10.1007/s11192-018-2958-5.

Haase, Marco; Seiler Zimmermann, Yvonne; Zimmermann, Heinz (2016): The impact of speculation on commodity futures markets – A review of the findings of 100 empirical

studies. In *Journal of Commodity Markets* 3 (1), pp. 1–15. DOI: 10.1016/j.jcomm.2016.07.006.

Haase, Marco; Seiler Zimmermann, Yvonne; Zimmermann, Heinz (2018): Permanent and transitory price shocks in commodity futures markets and their relation to speculation. In *Empir Econ* 56 (4), pp. 1359–1382. DOI: 10.1007/s00181-017-1387-2.

Hannesson, Rögnvaldur (2012): Does speculation drive the price of oil? In *OPEC Energy Review* 36 (2), pp. 125–137. DOI: 10.1111/j.1753-0237.2011.00207.x.

Harrer, Mathias; Cuijpers, Pim; Furukawa, Toshi; Ebert, David (2019): Doing Meta-Analysis in R: A Hands-on Guide. Available online at https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R.

Head, Megan L.; Holman, Luke; Lanfear, Rob; Kahn, Andrew T.; Jennions, Michael D. (2015): The extent and consequences of p-hacking in science. In *PLoS biology* 13 (3), e1002106. DOI: 10.1371/journal.pbio.1002106.

Hernandez, Manuel; Torero, Maximo (2010): Examining the Dynamic Relationship between Spot and Future Prices of Agricultural Commodities. discussion paper. Edited by International Food Policy Research Institute.

Hicks, J. R. (1946): Value and capital. 2. ed. Oxford U.P.: Clarendon Pr.

Hiemstra, Craig; Jones, Jonathan D. (1994): Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. In *The Journal of Finance* 49 (5), pp. 1639–1664. DOI: 10.1111/j.1540-6261.1994.tb04776.x.

Hirshleifer, J. (1977): The Theory of Speculation Under Alternative Regimes of Markets. In *The Journal of Finance* 32 (4), p. 975. DOI: 10.2307/2326507.

Huchet, Nicolas; Fam, Papa Gueye (2016): The role of speculation in international futures markets on commodity prices. In *Research in International Business and Finance* 37, pp. 49–65. DOI: 10.1016/j.ribaf.2015.09.034.

Hull, John (2006): Optionen, Futures und andere Derivate. 6. Aufl. München: Pearson Studium (Wi. Wirtschaft).

Iglewicz, Boris; Hoaglin, David Caster (1993): How to detect and handle outliers. In *Asq Press* (16).

Ioannidis, John P. A. (2005): Why most published research findings are false. In *PLoS medicine* 2 (8), e124. DOI: 10.1371/journal.pmed.0020124.

Irwin, S. H.; Garcia, P.; Good, D. L.; Kunda, E. L. (2011): Spreads and Non-Convergence in Chicago Board of Trade Corn, Soybean, and Wheat Futures: Are Index Funds to Blame? In *Journal of Economic Surveys* 33 (1), pp. 116–142. DOI: 10.1093/aepp/ppr001.

Irwin, Scott H.; Sanders, Dwight R. (2010): The Impact of Index and Swap Funds on Commodity Futures Markets. Selected studies on various food, agriculture and fisheries issues from the OECD Trade and Agriculture Directorate. working paper. OECD Food, Agriculture and Fisheries Papers (Selected studies on various food, agriculture and fisheries issues from the OECD Trade and Agriculture Directorate.). Available online at https://www.oecd-ilibrary.org/agriculture-and-food/the-impact-of-index-and-swap-funds-on-commodity-futures-markets_5kmd40wl1t5f-en.

Irwin, Scott H.; Sanders, Dwight R. (2012): Testing the Masters Hypothesis in commodity futures markets. In *Energy Economics* 34 (1), pp. 256–269. DOI: 10.1016/j.eneco.2011.10.008.

Irwin, Scott H.; Sanders, Dwight R.; Merrin, Robert P. (2009): Devil or Angel? The Role of Speculation in the Recent Commodity Price Boom (and Bust). In *J. Agric. Appl. Econ.* 41 (2), pp. 377–391. DOI: 10.1017/s1074070800002856.

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2013): An Introduction to Statistical Learning. New York, NY: Springer New York (103).

Jeong, Kiho; Härdle, Wolfgang K.; Song, Song (2012): A consistent nonparametric test for causality in quantile. In *Econom. Theory* 28 (4), pp. 861–887. DOI: 10.1017/S0266466611000685.

Kaldor, N. (1976): Speculation and Economic Stability. In B. A. Goss, B. S. Yamey (Eds.): The Economics of Futures Trading. London: Palgrave Macmillan London UK, pp. 111–123.

Keynes, John Maynard (1930): Treatise on money. New York.

Koenker, Roger (1981): A note on studentizing a test for heteroscedasticity. In *Journal of Econometrics* 17 (1), pp. 107–112. DOI: 10.1016/0304-4076(81)90062-2.

Koenker, Roger; Bassett, Gilbert (1978): Regression Quantiles. In *Econometrica* 46 (1), p. 33. DOI: 10.2307/1913643.

Koenker, Roger; Machado, Jose A. F. (1999): Goodness of Fit and Related Inference Processes for Quantile Regression. In *Journal of the American Statistical Association* 94 (448), pp. 1296–1310.

Kyle, Albert S. (1985): Continuous Auctions and Insider Trading. In *Econometrica* 53 (6), p. 1315. DOI: 10.2307/1913210.

Lee, Tae-Hwy; Yang, Weiping (2012): Money–Income Granger-Causality in Quantiles. In Dek Terrell, Daniel Millimet (Eds.): 30th Anniversary Edition, vol. 30: Emerald Group Publishing Limited (Advances in Econometrics), pp. 385–409.

Lescaroux, François (2009): On the excess co-movement of commodity prices—A note about the role of fundamental factors in short-run dynamics. In *Energy Policy* 37 (10), pp. 3906–3913. DOI: 10.1016/j.enpol.2009.05.013.

Leys, Christophe; Klein, Olivier; Dominicy, Yves; Ley, Christophe (2018): Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. In *Journal of Experimental Social Psychology* 74, pp. 150–156. DOI: 10.1016/j.jesp.2017.09.011.

Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (2008): Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

Lüdecke, Daniel (2020): Outliers detection (check for influential observations). Edited by RDocumentation. Available online at https://www.rdocumentation.org/packages/performance/versions/0.4.4/topics/check_outliers, checked on 3/17/2020.

Lütkepohl, Helmut (2007): New introduction to multiple time series analysis. Corr. 2nd print. Berlin, Heidelberg: Springer.

MacKinnon, James G.; White, Halbert (1985): Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. In *Journal of Econometrics* 29 (3), pp. 305–325. DOI: 10.1016/0304-4076(85)90158-7.

Malkiel, Burton G.; Fama, Eugene F. (1970): Efficient capital markets: A review of theory and empirical work. In *The Journal of Finance* 25 (2), pp. 383–417. DOI: 10.1111/j.1540-6261.1970.tb00518.x.

Malliaris, A. G.; Urrutia, Jorge L. (1998): Volume and price relationships: Hypotheses and testing for agricultural futures. In *J. Fut. Mark.* 18 (1), pp. 53–72. DOI: 10.1002/(sici)1096-9934(199802)18:1%3C53::aid-fut3%3E3.0.co;2-a.

Massot, Albert; Azevedo, Filipa; Ragonnaud, Guillaume (2013): Research on: Regulating Agricultural Derivatives Markets. Edited by European Commission. Available online at https://www.europarl.europa.eu/thinktank/de/document.html?reference=IPOL-AGRI_DV%282013%29513989, checked on 3/28/2020.

Mathur, Kritika; Kaicker, Nidhi; Gaiha, Raghav; Imai, Katsushi S.; Thapa, Ganesh (2013): Financialisation of food commodity markets, price surge and volatility: new evidence. discussion paper, pp. 149–176. DOI: 10.4337/9781781004296.00013.

Mayer, Herbert; Rathgeber, Andreas; Wanner, Markus (2017): Financialization of metal markets: Does futures trading influence spot prices and volatility? In *Resources Policy* 53, pp. 300–316. DOI: 10.1016/j.resourpol.2017.06.011.

Mayer, Jörg (2012): The Growing Financialisation of Commodity Markets: Divergences between Index Investors and Money Managers. In *Journal of Development Studies* 48 (6), pp. 751–767. DOI: 10.1080/00220388.2011.649261.

Meijerink, G. W.; Shutes, K.; Herder, A.; van Gelder, J. W. (2012): Food prices and speculation on agricultural futures markets: literature survey and interviews. In *Rapport - Landbouw-Economisch Instituut* (No.2012-009).

Merino, Antonio; Albacete, Rebeca (2010): Econometric modelling for short-term oil price forecasting. In *OPEC Energy Review* 34 (1), pp. 25–41. DOI: 10.1111/j.1753-0237.2010.00171.x.

Mukherjee, Kedar Nath (2011): Impact of Futures Trading on Indian Agricultural Commodity Market. working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.1763910.

Naderian, Mohammad Amin; Javan, Afshin (2017): Distortionary effect of trading activity in NYMEX crude oil futures market: post crisis. In *OPEC Energy Rev* 41 (1), pp. 23–44. DOI: 10.1111/opec.12092.

Natanelov, Valeri; Alam, Mohammad J.; McKenzie, Andrew M.; van Huylenbroeck, Guido (2011): Is there co-movement of agricultural commodities futures prices and crude oil? In *Energy Policy* 39 (9), pp. 4971–4984. DOI: 10.1016/j.enpol.2011.06.016.

Nath, Golaka C.; Lingareddy, Tulsi (2008): Impact of futures trading on commodity prices. In *Economic and Political Weekly* 2008, pp. 18–23.

Obadi, Saleh Mothana; Korcek, Matej (2018): The Crude Oil Price and Speculations: Investigation Using Granger Causality Test. In *International Journal of Energy Economics and Policy* 2018 (8), pp. 275–282. Available online at https://econpapers.repec.org/article/ecojourn2/2018-03-32.htm.

Often, Einar M.; Wisen, Craig H. (2013): Disaggregated Commitment Of Traders Data And Prospective Price Effects. In *JABR* 29 (5), p. 1381. DOI: 10.19030/jabr.v29i5.8021.

Peri, Massimo; Baldi, Lucia; Vandone, Daniela (2013): Price discovery in commodity markets. In *Applied Economics Letters* 20 (4), pp. 397–403. DOI: 10.1080/13504851.2012.709590.

Petersen, Mitchell A. (2009): Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. In *Rev. Financ. Stud.* 22 (1), pp. 435–480. DOI: 10.1093/rfs/hhn053.

Pindyck, Robert; Rotemberg, Julio (1988): The Excess Co-Movement of Commodity Prices. Cambridge, MA.

Pradhananga, Manisha (2014): Financialization of the commodities futures markets and its effects on prices. Doctoral Dissertation.

Prokopczuk, Marcel; Symeonidis, Lazaros; Verlaat, Timo (2014): Rising and Volatile Food Prices: Are Index Fund Investors to Blame? working paper. In *SSRN Journal*. DOI: 10.2139/ssrn.2450397.

Sanders, Dwight R.; Boris, Keith; Manfredo, Mark (2004): Hedgers, funds, and small speculators in the energy futures markets: an analysis of the CFTC's Commitments of Traders reports. In *Energy Economics* 26 (3), pp. 425–445. DOI: 10.1016/j.eneco.2004.04.010.

Sanders, Dwight R.; Irwin, Scott H. (2011a): New Evidence on the Impact of Index Funds in U.S. Grain Futures Markets. In *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 59 (4), pp. 519–532. DOI: 10.1111/j.1744-7976.2011.01226.x.

Sanders, Dwight R.; Irwin, Scott H. (2011b): The Impact of Index Funds in Commodity Futures Markets: A Systems Approach. In *JAI* 14 (1), pp. 40–49. DOI: 10.3905/jai.2011.14.1.040.

Sanders, Dwight R.; Irwin, Scott H. (2014): Energy futures prices and commodity index investment: New evidence from firm-level position data. In *Energy Economics* 46, S57-S68. DOI: 10.1016/j.eneco.2014.09.005.

Sanders, Dwight R.; Irwin, Scott H.; Merrin, Robert P. (2009): Smart Money: The Forecasting Ability of CFTC Large Traders in Agricultural Futures Markets. In *Journal of Agricultural and Resource Economics* 2009 (34).

Sassi, Maria; Werner, Harald A. (2013): Non-commercial actors and the recent futures prices of wheat. In *Economia & Diritto Agroalimentare* (17), pp. 309–330.

Schwarz, Gideon (1978): Estimating the Dimension of a Model. In *Ann. Statist.* 6 (2), pp. 461–464. DOI: 10.1214/aos/1176344136.

Sehgal, Sanjay; Rajput, Namita; Dua, Rajeev Kumar (2012): Futures Trading and Spot Market Volatility: Evidence from Indian Commodity Markets. In *AJFA* 4 (2). DOI: 10.5296/ajfa.v4i2.1990.

Sen, Sunanda; Paul, Mahua (2010): Trading in India's commodity future markets. working paper. Institute for Studies in Industrial Development.

Serrao, Amílcar (2016): A controversial debate between financial speculation and changes in agricultural commodity spot prices. working paper. Evora University.

Shanker, Latha (2017): New indices of adequate and excess speculation and their relationship with volatility in the crude oil futures market. In *Journal of Commodity Markets* 5, pp. 18–35. DOI: 10.1016/j.jcomm.2016.11.003.

Shanmugam, Velmurugan; Armah, Paul (2012): Role of speculators in agricultural commodity price spikes during 2006-2011. In *Academy of accounting and financial studies journal* 2012 (16), pp. 97–114.

Sharma, Dinesh Kumar; Malhotra, Meenakshi (2015): Impact of futures trading on volatility of spot market-a case of guar seed. In *Agricultural Finance Review* 75 (3), pp. 416–431. DOI: 10.1108/AFR-03-2014-0005.

Sharma, Tanushree (2016): The Impact of Future Trading on Volatility in Agriculture Commodity: A Case of Pepper. In *The IUP Journal of Financial Risk Management*, pp. 47–55.

Song, Xiaojun; Taamouti, Abderrahim (2020): Measuring Granger Causality in Quantiles*. In *Journal of Business & Economic Statistics*, pp. 1–48. DOI: 10.1080/07350015.2020.1739531.

Stanley, T. D. (2007): Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection. In *Oxford Bull Econ & Stats* 0 (0), 070921170652004-??? DOI: 10.1111/j.1468-0084.2007.00487.x.

Stanley, T. D.; Doucouliagos, Hristos (2016): Meta-regression analysis in economics and business. First issued in paperback 2016. London, New York: Routledge Taylor & Francis Group (Routledge Advances in Research Methods, 5).

Stanley, T. D.; Doucouliagos, Hristos (2017): Neither fixed nor random: weighted least squares meta-regression. In *Research synthesis methods* 8 (1), pp. 19–42. DOI: 10.1002/jrsm.1211.

Stanley, T. D.; Doucouliagos, Hristos; Giles, Margaret; Heckemeyer, Jost H.; Johnston, Robert J.; Laroche, Patrice et al. (2013): Meta-Analysis of economics research reporting guidelines. In *J Economic Surveys* 27 (2), pp. 390–394. DOI: 10.1111/joes.12008.

Stanley, T. D.; Jarrell, Stephen B. (2005): Meta-Regression Analysis: A Quantitative Method of Literature Surveys. In *J Economic Surveys* 19 (3), pp. 299–308. DOI: 10.1111/j.0950-0804.2005.00249.x.

Stanley, T.D (2001): Wheat From Chaff: Meta-Analysis As Quantitative Literature Review. In *Journal of Economic Perspectives* 15 (3), pp. 131–150. DOI: 10.1257/jep.15.3.131.

Steiner, Manfred; Bruns, Christoph (2007): Wertpapiermanagement. Professionelle Wertpapieranalyse und Portfoliostrukturierung. 9. Aufl. Stuttgart: Schäffer-Poeschel Verlag für Wirtschaft Steuern Recht GmbH (Handelsblatt-Bücher).

Stock, James; Sims, C.; Watson, M. (1990): Inference in Linear Time Series Models with Some Unit Roots. In *Econometrica* 58 (1), pp. 113–144.

Stoll, Hans R.; Whaley, Robert E. (2010): Commodity Index Investing: Speculation or Diversification? In *SSRN Journal*. DOI: 10.2139/ssrn.1633908.

Su, Liangjun; White, Halbert (2008): A nonparametric Hellinger metric test for conditional independence. In *Econom. Theory* 24 (4), pp. 829–864. DOI: 10.1017/S0266466608080341.

Telser, Lester G. (1959): A Theory of Speculation Relating Profitability and Stability. In *The Review of Economics and Statistics* 41 (3), p. 295. DOI: 10.2307/1927455.

Telser, Lester G. (1967): The supply of speculative services in wheat, corn, and soybeans. In *Food Research Institute Studies* 7, pp. 131–176.

Thompson, Simon G.; Higgins, Julian P. T. (2002): How should meta-regression analyses be undertaken and interpreted? In *Statistics in medicine* 21 (11), pp. 1559–1573. DOI: 10.1002/sim.1187.

Toda, Hiro Y.; Phillips, Peter C. B. (1993): Vector Autoregressions and Causality. In *Econometrica* 61 (6), p. 1367. DOI: 10.2307/2951647.

Toda, Hiro Y.; Yamamoto, Taku (1995): Statistical inference in vector autoregressions with possibly integrated processes. In *Journal of Econometrics* 66 (1-2), pp. 225–250. DOI: 10.1016/0304-4076(94)01616-8.

Troster, Victor (2018): Testing for Granger-causality in quantiles. In *Econometric Reviews* 37 (8), pp. 850–866. DOI: 10.1080/07474938.2016.1172400.

Tse, Yiuman; Williams, Michael R. (2013): Does Index Speculation Impact Commodity Prices? An Intraday Analysis. In *Financial Review* 48 (3), pp. 365–383. DOI: 10.1111/fire.12007.

UNCTAD (2011): Price Volatility in Food and Agricultural Markets: Policy Responses. Policy Report including contributions by FAO, IFAD, IMF,OECD, UNCTAD, WFP, the World Bank, the WTO, IFPRI and the UN HLTF, pp. 1–68.

Will, Matthias Georg; Prehn, Sören; Pies, Ingo; Glauben, Thomas (2012): Is financial speculation with agricultural commodities harmful or helpful? A literature review of current empirical research. discussion paper. Halle, Halle, Saale: Martin-Luther-Univ. Halle-Wittenberg, Lehrstuhl für Wirtschaftsethik; Universitäts- und Landesbibliothek Sachsen-Anhalt (2012,27).

Wimmer, Thomas; Geyer-Klingeberg, Jerome; Hütter, Marie; Schmid, Florian; Rathgeber, Andreas (2020): The Impact of Speculation on Commodity Prices: A Meta-Granger Analysis. Universität Augsburg. Researchgate.

Wooldridge, Jeffrey M. (2003): Cluster-Sample Methods in Applied Econometrics. In *American Economic Review* 93 (2), pp. 133–138. DOI: 10.1257/000282803321946930.

Working, Holbrook (1953): Futures trading and hedging. In *The American Economic Review*, pp. 314–343.

Working, Holbrook (1961): New Concepts Concerning Futures Markets and Prices. Papers and Proceedings of the Seventy-Third Annual Meeting of the American. In *The American Economic Review* (51), pp. 160–163.

Yang, Jian; Balyeat, R. Brian; Leatham, David J. (2005): Futures Trading Activity and Commodity Cash Price Volatility. In *Journal of Business Finance & Accounting* 32 (1-2), pp. 297–323. DOI: 10.1111/j.0306-686x.2005.00595.x.

Zapata, Hector O.; Rambaldi, Alicia N. (1997): Monte Carlo Evidence on Cointegration and Causation. In *Oxford Bull Econ & Stats* 59 (2), pp. 285–298. DOI: 10.1111/1468-0084.00065.

Zeileis, Achim (2004): Econometric Computing with HC and HAC Covariance Matrix Estimators. In *J. Stat. Soft.* 11 (10). DOI: 10.18637/jss.v011.i10.

Zimmermann, Christian; Barrueco Cruz, José Manuel; Baum, Kit; Karlsson, Sune; Krichel, Thomas; Klink, Markus (2020): Research Papers in Economics IDEAS. Edited by Federal Reserve Bank of St. Louis. Available online at https://ideas.repec.org/, checked on 4/19/2020.

Appendix A: Operation and Calculation

Table 8: Calculation of sample sizes for different periods

Period	Calculation
Daily	$(Day_{end} - Day_{start})$
Weekly	$(Day_{end} - Day_{start})/5$
Monthly	$(Year_{end} - Year_{start}) * 12 + Month_{start} - Month_{end}$
Quarterly	*Manually calculated*

Note: Difference of days is calculated in Excel with the WORKDAY() command

Table 9: Calculation of p-values

Used statistic	Used assumption
F-statistic	Probability of the right-sided F-distribution
Chi²-statistic	a. Transformation into F-statistics (Chi²-statistic /No. of lags) b. Probability of the right-sided F-distribution
z-statistic	Probability mass function of the standard normal distribution

Note: Probabilities calculated in Excel with the F.VERT.RE() and the NORM.S.VERT() command.

Appendix B: List of analyzed empirical studies for subset S2M

Table 10: List of analyzed empirical studies for subset S2M

Study ID	Autor	No. Of obs.	Reason for exclusion
1	Abdullahi et al. 2014	4	
2	Algieri 2016	57	
3	Alquist and Gervais 2013	Excluded	No lags indicated
4	Amann et al. 2013	Excluded	No lags indicated
5	Antonakakis et al. 2018	Excluded	No commodity indicated
6	Aulerich et al. 2010	Excluded	No lags indicated
7	Aulerich et al. 2014	72	
8	Babalos and Balcilar 2017	Excluded	No commodity indicated
9	Baldi et al. 2011	Excluded	No speculation measurement, but future – spot market influence
10	Bohl et al. 2018	16	
11	Borin and Di Nino 2012	9	
12	Bos and van der Molen 2012	Excluded	No speculation measurement, but future -real world indicator influence
13	Brunetti and Buyuksahin 2009	36	
14	Brunetti et al. 2011	18	
15	Brunetti et al. 2013	Excluded	No lags indicated
16	Bu 2011	8	
17	Buyuksahin and Harris 2011	374	
18	Capelle-Blancard and Coulibaly 2012	36	
19	Chakraborty and Das 2013	44	
20	Ciner 2002	3	
21	Coleman and Dark 2012	18	
22	Ding et al. 2014	12	
23	Ederer et al. 2013	10	
24	Etienne et al. 2017	16	
25	Fagan and Gencay 2008	1	
26	Fujihara and Mougou 1997	3	
27	Ghalayini 2011	2	
28	Ghalayini 2017	Excluded	No speculation measurement, but future – spot market influence
29	Gilbert and Pfuderer 2012	20	

Study ID	Autor	No. Of obs.	Reason for exclusion
30	Gilbert and Pfuderer 2014	15	
31	Gilbert 2009	14	
32	Gilbert 2010a	16	
33	Gilbert 2010b	9	
34	Girardi 2015	Excluded	No speculation measure, but index trading measure
35	Gupta et al. 2018	16	
36	Haase et al. 2018	18	
37	Hannesson 2012	Excluded	No speculation measurement, but future – spot market influence
38	Hernandez and Torero 2010	Excluded	No speculation measurement, but future – spot market influence
39	Huchet and Fam 2016	Excluded	No lags indicated
40	Irwin and Sanders 2010	38	
41	Irwin and Sanders 2012	18	
42	Irwin et al. 2009	10	
43	Irwin et al. 2011	12	
44	Malliaris and Urrutia 1998	18	
45	Mathur et al. 2013	Excluded	No commodity indicated
46	Mayer et al. 2017	Excluded	No lags indicated
47	Mayer 2012	48	
48	Merino and Albacete 2010	Excluded	No lags indicated
49	Mukherjee 2011	Excluded	No speculation measurement, but future – spot market influence
50	Naderian and Javan 2017	21	
51	Obadi and Korcek 2018	3	
52	Often and Wisen 2013	Excluded	No lags indicated
53	Peri et al. 2013	Excluded	No speculation measurement, but future – spot market influence
54	Prokopczuk et al. 2014	144	
55	Sanders and Irwin 2011a	32	
56	Sanders and Irwin 2011b	44	
57	Sanders and Irwin 2014	8	
58	Sanders et al. 2004	4	
59	Sanders et al. 2009	40	
60	Sassi and Werner 2013	4	

Study ID	Autor	No. Of obs.	Reason for exclusion
61	Sehgal et al. 2012	Excluded	No lags indicated
62	Sen and Paul 2010	Excluded	No speculation measurement, but future – spot market influence
63	Serrao 2016	Excluded	No lags & sample size indicated
64	Shanker 2017	4	
65	Shanmugam and Armah 2012	Excluded	No lags indicated
66	Sharma and Malhotra 2015	2	
67	Sharma 2016	Excluded	No lags indicated
68	Stoll and Whaley 2010	12	
69	Tse and Williams 2013	Excluded	No speculation measurement, but co-movement measurement
70	Yang et al. 2005	28	

Appendix C: Multiple MRA, including variance calculation following Wimmer et al. (2020)

Table 11: Multiple MRA, including variance calculation approach following Wimmer et al. (2020)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
	-	$\sqrt{DF_i}$	$\frac{1}{obs.}$	studyquality	$\frac{1}{variance}$
Subset A					
Intercept	3.053	3.230	1.575	2.757	0.547
	(1.6630)	(1.8311)	(1.4044)	(1.3470)	(0.3588)
$\sqrt{DF_i}$	0.043	0.061	-0.006	0.051	0.027
	(1.4402)	(1.8612)	(-0.2007)	(1.5886)	(1.3158)
Lags m:	0.135	0.143	0.234	0.129	0.280
	(1.0731)	(1.0374)	(1.9081)	(1.0294)	(1.3000)
Commodity groups					
Energy			Baseline		
Metal	**-3.037**	**-3.985**	**-2.880**	**-2.602**	**-2.161**
(small sample)	**(-2.4299)***	**(-2.8136)****	**(-2.6832)****	**(-2.1814)***	**(-2.3089)***
Soft Commodities	**-1.787**	**-1.752**	-0.143	**-1.612**	-0.811
	(-2.4044)*	**(-2.3297)***	(-0.2142)	**(-2.2909)***	(-1.8633)
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	0.086	0.090	0.376	0.152	0.308
	(0.1110)	(0.0892)	(0.5748)	(0.1684)	(0.4707)
t-statistic	1.057	1.269	0.385	1.209	0.805
	(0.7755)	(1.0382)	(0.5010)	(0.7802)	(1.4744)
Differences	0.325	0.247	-0.253	0.360	0.292
	(0.5131)	(0.3954)	(-0.5232)	(0.4909)	(0.6030)
Log x var	-0.828	-0.635	**-1.076**	-0.793	**-0.595**
	(-1.8850)	(-0.9726)	**(-2.3105)***	(-1.7381)	**(-2.5709)***
Sum	0.971	1.370	-1.041	0.841	**0.040**
	(0.9084)	(1.0107)	(-1.1802)	(0.7410)	**(0.0668)**

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Var/Vec	-0.669	-0.599	0.014	-0.730	0.429
	(-1.0947)	(-0.8351)	(0.0241)	(-0.9652)	(0.8275)
z-variable	-0.365	-0.566	0.876	-0.047	-0.153
	(-0.5044)	(-0.7072)	(1.0963)	(-0.0762)	(-0.3776)
Linear GC			Baseline		
non-parametric GC	-1.892	-2.502	-0.955	-2.076	-1.431
	(-1.0828)	(-1.4487)	(-0.8504)	(-1.1009)	(-1.8753)
Multivariate GC	1.412	1.303	0.034	1.514	0.669
	(1.2240)	(1.2103)	(0.0496)	(1.1696)	(1.2258)
GC in quantile			Excluded due to (multi-) collinearity		
AIC plus	0.686	0.782	-0.095	0.974	0.182
	(0.7822)	(0.7297)	(-0.2171)	(0.8934)	(0.3846)
CFTC	-1.791	-2.026	-0.664	**-2.098**	-0.638
	(-1.9413)	(-1.7602)	(-1.0279)	**(-1.9707)***	(-1.0083)
Data time characteristics					
Weekly			Excluded due to (multi-) collinearity		
Daily			Excluded due to (multi-) collinearity		
Monthly	0.616	0.874	**0.938**	0.394	0.105
	(1.0729)	(1.1929)	**(1.9701)***	(0.6181)	(0.2417)
quarterly			Excluded due to (multi-) collinearity		
After 2007	0.107	-0.112	0.459	0.159	-0.156
	(0.1831)	(-0.1438)	(1.1002)	(0.2475)	(-0.3897)
Publication characteristics					
Influenced	-0.821	-1.388	-0.272	-0.751	0.024
	(-1.0075)	(-1.5886)	(-0.5042)	(-0.8468)	(0.0386)
Ranking	-0.120	-0.829	-0.407	-0.017	0.600
	(-0.0959)	(-0.6114)	(-0.6253)	(-0.0122)	(0.7664)
Proxy variable characteristics					
Volatility	-0.610	-0.619	-0.829	-0.657	-0.252
	(-1.2251)	(-1.2341)	(-1.5375)	(-1.2074)	(-0.5552)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
OI	-1.080 (-1.1206)	-1.371 (-1.2627)	-0.474 (-0.5956)	-1.172 (-1.1263)	-0.395 (-0.6302)
Volume	-1.597 (-1.1395)	-2.203 (-1.3940)	0.200 (0.1773)	-1.787 (-1.1667)	-0.403 (0.4338)
Studentized Breusch-Pagan	1.688e-09	1.229e-09	1.132e-0.08	1.103e-09	1.502e-08
#Studies	45	44	45	45	45
#Observations	1,501	1,485	1,467	1,496	1,486
Subset B					
Intercept	**2.192** **(2.1663)***	**2.198** **(2.2897)***	1.214 (1.1368)	2.251 (2.2330)*	0.721 (0.6288)
$\sqrt{DF_i}$	0.020 (0.9132)	0.020 (0.9358)	-0.008 (-0.3614)	0.019 (0.8346)	0.018 (1.0062)
Lags m:	0.070 (1.2060)	0.103 (1.3480)	**0.154** **(2.7217)****	0.066 (1.1265)	0.139 (1.3154)
Commodity groups					
Energy			Baseline		
Metal (small sample)	**-1.408** **(-3.5981)*****	**-1.683** **(-3.1358)****	**-1.430** **(-3.6228)*****	**-1.151** **(-2.7125)****	**-1.458** **(-2.9280)****
Soft Commodities	-0.443 (-1.3014)	-0.207 (-0.4305)	0.379 (1.2273)	-0.416 (-1.1914)	-0.346 (-1.4850)
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	-0.318 (-0.6059)	0.044 (0.0715)	0.612 (1.2275)	-0.465 (-0.7748)	0.003 (0.0053)
t-statistic	0.026 (0.0442)	0.060 (0.1013	-0.461 (-0.905)	0.066 (0.1092)	0.321 (0.7472)
Differences	-0.405 (-1.3252)	-0.419 (-1.1817)	**-0.874** **(-2.0326)***	-0.405 (-1.2002)	-0.165 (-0.4704)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Log x var	**-0.806**	**-0.909**	-0.700	**-0.777**	**-0.652**
	(-3.2002)**	**(-3.6347)*****	(-1.9466)	**(-3.2331)****	**(-2.9436)****
Sum	-0.237	-0.176	-0.533	-0.271	-0.603
	(-0.3796)	(-0.2423)	(-0.8840)	(-0.4235)	(-1.0682)
Var/Vec	-0.006	-0.129	-0.582	0.164	0.639
	(-0.0134)	(-0.2462)	(-1.3098)	(0.3385)	(1.4622)
Z-variable	-0.063	0.231	0.612	-0.119	-0.149
	(-0.1229)	(0.4757)	(0.7417)	(-0.2314)	(-0.3219)
Linear GC			Baseline		
Non-parametric GC	0.207	0.570	0.956	0.177	-0.312
	(0.3414)	(0.9838)	(1.9433)	(0.2948)	(-0.7530)
Multivariate GC	-0.755	-0.744	-1.355	-0.972	-0.777
	(-1.1867)	(-1.1237)	(-1.9515)	(-1.4412)	(-1.7184)
GC in quantile			Excluded due to (multi-) collinearity		
AIC plus	-0.683	**-0.941**	-0.698	-0.767	-0.397
	(-1.6420)	**(-2.0830)***	(-1.5809)	(-1.7326)	(-1.0725)
CFTC	-0.721	-0.686	0.088	-0.836	-0.315
	(-1.1280)	(-1.0799)	(0.1705)	(-1.2824)	(-0.4687)
Data time characteristics					
Weekly			Excluded due to (multi-) collinearity		
Daily			Excluded due to (multi-) collinearity		
Monthly	0.528	0.761	0.730	0.431	0.063
	(0.8171)	(1.0308)	(1.8870)	(0.6310)	(0.1051)
Quarterly			Excluded due to (multi-) collinearity		
After 2007	0.181	0.143	0.346	0.189	-0.011
	(0.4226)	(0.2921)	(1.1726)	(0.4314)	(-0.0293)
Publication characteristics					
Influenced	-0.433	-0.818	-0.114	-0.478	0.090
	(-0.9747)	(-1.7811)	(-0.3939)	(-1.0369)	(0.1890)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Ranking	0.808	0.562	0.308	0.792	0.855
	(1.4185)	(0.9170)	(0.7197)	(1.4138)	(1.6323)
Proxy variable characteristics					
Volatility	-0.293	-0.454	0.002	-0.307	-0.215
	(-0.7695)	(-1.0741)	(0.0065)	(-0.7825)	(-0.4502)
OI	**-1.317**	**-1.601**	-0.527	**-1.377**	-0.603
	(-2.6339)**	**(-3.1594)****	(-1.0227)	**(-2.6026)****	(-1.1206)
Volume	-0.852	-0.890	0.733	-0.833	-0.130
	(-0.9486)	(-0.9343)	(0.7040)	(-0.8740)	(-0.1614)
Studentized Breusch-Pagan	2.2e-16	8.285e-16	2.6e-16	2.2e-16	2.2e-16
#Studies	44	44	44	44	43
#Observations	1,444	1,416	1,405	1,445	1,396

Appendix D: Multiple MRA, using experimental cluster-robust standard errors and including variance calculation following Wimmer et al. (2020)

Table 12: Multiple MRA, using a experimental cluster-robust standard error approach and including variance calculation approach following Wimmer et al. (2020)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
	-	$\sqrt{DF_i}$	$\frac{1}{obs.}$	studyquality	$\frac{1}{variance}$
Subset A					
Intercept	2.655	3.328	1.396	2.323	-0.080
	(1.3275)	(1.7938)	(1.0533)	(1.0492)	(-0.0489)
$\sqrt{DF_i}$	**0.089**	**0.107**	0.032	**0.092**	**0.075**
	(2.6740)**	**(4.0082)*****	(0.9304)	**(2.6887)****	**(3.0801)****
Lags m:	0.133	0.143	0.195	0.123	0.292
	(1.1573)	(1.1745)	(1.6297)	(1.0841)	(1.3982)
Commodity groups					
Energy			Baseline		
Metal	**-3.370**	**-4.326**	**-2.940**	**-2.976**	**-2.600**
(small sample)	**(-2.8547)****	**(-3.4737)*****	**(-3.1567)****	**(-2.7024)****	**(-2.9119)****
Soft Commodities	**-1.996**	**-2.052**	-0.651	**-1.800**	**-1.017**
	(-2.2768)*	**(-2.0360)***	(-0.9661)	**(-2.1051)***	**(-2.2732)***
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	0.077	0.131	0.251	0.147	0.595
	(0.0861)	(0.1180)	(0.3590)	(0.1452)	(0.7626)
t-statistic	0.842	0.680	-0.002	0.976	0.454
	(0.6627)	(0.5500)	(-0.0020)	(0.7098)	(0.8793)
Differences	0.137	-0.253	-0.413	0.184	0.098
	(0.2181)	(-0.4159)	(-0.9194)	(0.2562)	(0.1714)
Log x var	-0.668	-0.724	**-1.064**	-0.473	**-0.508**
	(-1.2235)	(-0.9100)	**(-2.1179)***	(-0.8393)	**(-1.5242)**

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Sum	1.169	1.476	-1.018	1.051	0.573
	(1.0634)	(1.2226)	(-1.0626)	(0.9091)	(0.9807)
Var/Vec	-0.802	-0.931	0.118	-0.774	0.089
	(-1.1423)	(-1.2048)	(0.2530)	(-0.9560)	(0.2247)
z-variable	-1.146	-1.257	0.729	-1.233	-1.229
	(-1.1238)	(-1.1503)	(0.5655)	(-1.2382)	(-1.4562)
Linear GC			Baseline		
non-parametric GC	-1.926	-2.185	-1.040	-2.101	-1.332
	(-1.1674)	(-1.2507)	(-0.8411)	(-1.2006)	(-1.7922)
Multivariate GC	0.632	0.476	-0.363	0.612	-0.358
	(0.5599)	(0.4187)	(-0.4440)	(0.4889)	(-0.6167)
GC in quantile			Excluded due to (multi-) collinearity		
AIC plus	0.477	0.574	-0.007	0.680	0.026
	(0.5272)	(0.5334)	(-0.0206)	(0.6137)	(0.0582)
CFTC	-1.692	-1.826	-0.517	-1.878	-0.345
	(-1.8518)	(-1.7414)	(-0.7572)	(-1.7696)	(-0.4780)
Data time characteristics					
Weekly			Excluded due to (multi-) collinearity		
Daily			Excluded due to (multi-) collinearity		
Monthly	1.213	1.829	0.812	0.864	0.885
	(1.6271)	(2.0350)	(1.5265)	(1.1089)	(1.9420)
quarterly			Excluded due to (multi-) collinearity		
After 2007	0.359	0.167	0.501	0.397	0.134
	(0.5818)	(0.2064)	(1.0789)	(0.5931)	(0.3305)
Publication characteristics					
Influenced	-0.576	-1.051	-0.025	-0.473	0.207
	(-0.6780)	(-1.1861)	(-0.0371)	(-0.4995)	(0.3025)
Ranking	-0.752	-1.805	-0.593	-0.587	-0.015
	(-0.6232)	(-1.4011)	(-0.8328)	(-0.4319)	(-0.0174)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Proxy variable characteristics					
Volatility	-0.806 (-1.3528)	-1.077 (-1.6333)	-0.621 (-0.9150)	-0.747 (-1.2364)	-0.245 (-0.4797)
OI	-1.419 (-1.3744)	**-2.262** **(-2.1281)***	-0.352 (-0.4126)	-1.431 (-1.2863)	-0.598 (-0.7605)
Volume	-2.667 (-1.7973)	**-3.207** **(-2.1778)***	-1.181 (-0.9575)	-2.702 (-1.6867)	-1.126 (-1.0439)
Studentized Breusch-Pagan	1.624e-06	1.624e-06	1.624e-06	1.624e-06	1.624e-06
#Studies			46		
#Observations			1,560		
Subset B					
Intercept	**2.789** **(3.0163)****	**2.867** **(3.2114)****	1.141 (1.1117)	**2.702** **(2.8332)****	0.933 (0.8496)
$\sqrt{DF_i}$	0.029 (1.3129)	0.041 (1.8501)	0.015 (0.7411)	0.030 (1.3809)	0.034 (1.7147)
Lags m:	0.050 (1.1529)	0.055 (1.1235)	**0.084** **(1.8743)**	0.049 (1.0800)	0.087 (1.0409)
Commodity groups					
Energy			Baseline		
Metal (small sample)	**-1.586** **(-3.8730)*****	**-2.040** **(-3.4509)*****	**-1.737** **(-3.6434)*****	**-1.366** **(-2.9906)****	**-1.555** **(-2.9317)****
Soft Commodities	**-0.722** **(-2.6739)****	-0.655 (-1.8765)	-0.164 (-0.6545)	**-0.684** **(-2.5706)***	**-0.540** **(-2.4834)***
Data and test characteristics					
F-statistic			Baseline		
Chi²-statistic	-0.617 (-1.0507)	-0.504 (-0.7143)	0.148 (0.2684)	-0.686 (-1.0284)	0.084 (0.1270)
t-statistic	0.014 (0.0245)	0.345 (0.5927)	-0.331 (-0.7907)	0.114 (0.2114)	0.094 (0.2479)

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Differences	-0.484	-0.438	-0.595	-0.496	-0.344
	(-1.3605)	(-1.1254)	(-1.4399)	(-1.2857)	(-0.8832)
Log x var	**-0.877**	**-0.985**	**-0.643**	**-0.835**	**-0.753**
	(-3.5008)* **	**(-3.3642)*** **	**(-1.9703)***	**(-3.3947)*** **	**(-3.2591)** **
Sum	-0.244	-0.159	-0.307	-0.394	-0.282
	(-0.4853)	(-0.2690)	(-0.5762)	(-0.7819)	(-0.6941)
Var/Vec	-0.116	-0.077	-0.278	0.089	0.312
	(-0.4454)	(-0.2152)	(-0.7645)	(0.3157)	(1.2942)
Z-variable	-0.070	-0.135	0.365	-0.111	-0.192
	(-0.1275)	(-0.2665)	(0.5094)	(-0.2114)	(-0.4314)
Linear GC			Baseline		
Non-parametric GC	0.107	0.001	0.209	-0.021	-0.280
	(0.1801)	(0.0020)	(0.4084)	(-0.0372)	(-0.6860)
Multivariate GC	-0.835	-0.939	-1.179	-1.092	-0.855
	(-1.2036)	(-1.2578)	(-1.7441)	(-1.4716)	(-1.3863)
GC in quantile			Excluded due to (multi-) collinearity		
AIC plus	-0.550	-0.740	-0.066	-0.625	-0.312
	(-1.4251)	(-1.7169)	(-0.2046)	(-1.5146)	(-0.8751)
CFTC	-0.888	-0.853	0.213	-0.972	-0.235
	(-1.4008)	(-1.2830)	(0.4057)	(-1.4439)	(-0.3475)
Data time characteristics					
Weekly			Excluded due to (multi-) collinearity		
Daily			Excluded due to (multi-) collinearity		
Monthly	0.535	0.777	0.227	0.419	0.373
	(0.8925)	(1.0882)	(0.6275)	(0.6633)	(0.6789)
Quarterly			Excluded due to (multi-) collinearity		
After 2007	0.147	-0.011	0.153	0.144	0.079
	(0.3309)	(-0.0193)	(0.5348)	(0.3115)	(0.2116)
Publication characteristics					

	Base	WLS 1	WLS 2	WLS 3	WLS 4
Influenced	-0.628	-1.055	-0.285	-0.561	0.010
	(-1.4938)	(-2.5930)**	(-0.8406)	(-1.3050)	(0.0214)
Ranking	0.367	-0.166	-0.191	0.349	0.514
	(0.6606)	(-0.2529)	(-0.5087)	(0.5924)	(0.9560)
Proxy variable characteristics					
Volatility	-0.348	-0.599	0.314	-0.342	-0.002
	(-0.9383)	(-1.4859)	(0.6914)	(-0.9074)	(-0.0057)
OI	**-1.288**	**-1.568**	-0.005	**-1.215**	-0.558
	(-2.5118)*	**(-2.6921)****	(-0.0107)	**(-2.1597)***	(-1.0031)
Volume	-1.162	-1.276	0.160	-1.088	-0.446
	(-1.3694)	(-1.4636)	(0.1941)	(-1.2351)	(-0.5552)
Studentized Breusch-Pagan	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
#Studies			46		
#Observations			1,560		

Appendix E: Coding protocol

Study	Coding protocol
General assumptions	- Start and end dates, for dates where only annual and monthly information was given, are selected as 1st net working day of the month, ignoring regional and country specific holidays due to not proportionate complexity. - If period is daily: Only net working days were considered as trading days, ignoring potential OTC trading. Apart of that no adjustment for regional and country specific holidays are made due to not proportionate complexity. - Data from CFTC is assumed to be U.S. specific - "Standard GC" indicates the Granger causality test based on the Granger 1969 approach. I assumed that all "linear Granger causality tests" or "standard Granger causality tests" without further explanation or sources are also based on the Granger 1969 approach. - Lev/dif indication is always indicating for the x variable - CFTC data is automatically indicating US data - Func_form is assumed to be linear when no indication of log or ln or rel form is given - Hypothese_style is single for beta1 = beta2 = … = 0
Abdullahi et al. 2014	- Assumption: On page 432 they indicate unrestricted VAR in levels because returns and volume series are stationary I80) so lev is assumed - Assumption: Lag length indicated is valid for m and n
Algieri 2016	- Assumption: 1995:2 2012:1 is assumed to be 01.02.1995 till 02.01.2012 the same logic is applied for all the other dates. - Assumption: Lag length indicated is valid for m and n - Assumption: Hypothese_style is single - No date data for DP GC available, therefore dates of corresponding standard GC test are assumed (01.02.1995 – 02.01.2012)
Alquist and Gervais 2013	- Assumption: 1993M1, 2010M12, 2003M1, 2008M6, 2003M1 and 2010M12 means 1nd month of 1993 and so on. - Assumption: Data are first differences, because "changes" of variable x and y are used - Assumption: Hypothese_style is single - No specific lags were identified - Assumption: No system was identified, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
Amann et al. 2013	- <0.01 p values are simplified to 0.01 to be able to calculate with them - Assumption: Hypothese_style is single - No specific lags were indentified
Antonakakis et al. 2018	- Assumption: Data are first differences (return and flow) - Assumption: Lag length indicated is valid for m and n - Assumption: Hypothese_style is single
Aulerich et al. 2010	- Assumption: Clear indication of day and month for the start years 2006 and end year 2005 are not available. For the start year 2006 I assumed 02.01.2006 and for the end year 2005 I assumed 30.12.2005 - Assumption: Is "standard GC" because in Hamilton 1994 p. 302 they refer to Granger causality as proposed by Granger 1969 and popularized by Sims 1972 - Assumption: Data are dif for "change in CIT Net Long Open Interest" and lev for "CIT Net Long Open Interest as a Percent of Total Open Interest" - No specific lags were identified, but lags must be one of the following (1,1), (2,1), (1,2) - Assumption: No system was identified, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
Aulerich et al. 2014	- Assumption: Standard GC - Assumption: Data are in levels, because data are stationary

Study	Coding protocol
	- Data with Data-ID 13 till 24 omitted because the p-value is always the same. It is assumed to be a error of the author.
	- Assumption: No system was identified, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Data are in first differences (return and flow)
	- Assumption: Standard GC is used and not NGC, not 100% clear but indicated in the text
Babalos and	- Assumption: Lag length indicated is valid for m and n
Balcilar 2017	- Assumption: Hypothese_style is single
	- Calculated sample size o January 2004: Friday 02.01.2004 o September 2010: 01.09.2010 o December 2006: 01.12.2006
	- Qualifying date is deduced form figure 1: o 2007: 02.01.2007 o 2008: 02.10.2008 o 2005: 02.01.2005
	- Assumption: Hypothese_style is single
	- Assumption: Lag length consists of k + d
	- Assumption: Lag length indicated is valid for m and n
Baldi et al.	- Assumption: Some p-values are 0. Those p-value are calculated through conversion of Chi2 values to F values
2011	(Chi2/x lags) and conversion of F values through F.VERT.RE(F value, x lags, sample size) into p values
	- Sample size from paper
	- There is an irregularity: Sample size for Granger causality test for rapeseed oil with 738 is significantly smaller than sample size for contract with 2487
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Some p-values are 0. Those p-value are calculated from the F values through F.VERT.RE(F value,
Bohl et al. 2018	x lags, sample size) into p values
	- Assumption: Standard GC assumed; Wald test is expressed through Chi2 effect_typ
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Hypothese_style is single
Borin and Di	- Assumption: Some p-values are 0. Those p-value are calculated through conversion of Chi2 values to F values
Nino 2012	(Chi2/x lags) and conversion of F values through F.VERT.RE(F value, x lags, sample size) into p values
	- Assumption: Standard GC assumed
	- Assumption: Exact dates for period 1989 to 2008 are assumed to be 03.01.1989 and 31.12.2008
	- Calculated sample size on monthly base
	- Assumption: Indicated lag length of “up to 6 months” means lag of 6
Bos and van der	- Assumption: Lag length indicated is valid for m and n
Molen 2012	- Assumption: Hypothese_style is single
	- HF and FH both indicate “hedge funds”
	- Assumption: Page 31 Table 7 Panel B, not “Returns and Net Futures Positions in Levels” but “Returns and Net Futures Positions in First Difference”
	- Assumption: Page 33 Table 9, FH and HF are both variables for “Hedge Funds”
	- Calculated sample size on daily base
	- Rate of return and volatility are assumed to be “in levels”
	- Assumption: Lag length indicated is valid for m and n
Brunetti and	- Assumption: Hypothese_style is single
Buyuksahin	- Assumption: Lags were identified through several Wald tests and testing for significance, therefor the
2009	assumption is Wald test is expressed through Chi2 effect_typ

Study	Coding protocol
	- Assumption: Standard GC
	- Volatility is mainly based on trading position (see formula of page 16)
	- Calculated sample size on daily base
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Hypothese_style is single
Brunetti et al. 2011	- Assumption: Lags were identified through several Wald tests and testing for significance, therefor the assumption is Wald test is expressed through Chi2 effect_typ
	- Assumption: Standard GC
	- Calculated sample size on daily base
	- Assumption: herding metrics are first differences, as well as return and volatility data
Brunetti et al.	- Assumption: Hypothese_style is single
2013	- No specific lags were identified (only 0 lags for "system" variable is available)
	- Assumption Table 3: White Space is a comma
	- Calculated sample size on weekly base
	- Assumption: Hypothese_style is single
	- Cumulative impact indication is available but only indicated with significant / not significant and therefor omitted in the coding scheme
	- No specific lags were identified
	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC (also for instantaneous GC, because this is basically just a standard GC)
	- Assumption: Some p-values are 0. Those p-value for Chi2 statistics are calculated through conversion of Chi2 values to F values (Chi2/x lags) and conversion of F values through F.VERT.RE(F value, x lags, sample size) into p values. Those p-value for F-statistics are calculated from F values through F.VERT.RE(F value, x lags,
Bu 2011	sample size) into p values.
	- Assumption: Delta Day is lag
	- Assumption: Page 194 Table 7a Delta Day 1, Non-Commercials, Futures and Options, Position – Price, p-Wert: 0325 is 0,0325
	- Dolado and Lutkephol GC, reverse to modified Granger-causality test developed by Dolado and Lutkepohl (1996)
	- Analysis updates and enhances similar findings in the Interagency Task Force Interim Report on Crude Oil (ITF (2008)), so here is assumed that this Report influences the analysis
	- Calculated sample size on daily base
	- Assumption: Due to order of integration 1 first differences VARs are used. Even if position data are order of intergration 0, on page 189 it is indicated that first differenced VAR is not only used for price variables.
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Hypothese_style is single, because coefficients are "jointly significantly different from zero"
Buyuksahin and Harris 2011	- Assumption: Some p-values are 0. Those p-value are calculated through conversion of Chi2 values to F values (Chi2/x lags) and conversion of F values through F.VERT.RE(F value, x lags, sample size) into p values
	- Assumption: Exact dates for period Augsburg 2008 and September 2008 are assumed to be 29.08.2008 and 01.09.2008
	- Panel GC is the panel Granger causality testing approach by Kònya (2006), there for not coded, because data is not comparable
Capelle-Blancard and	- P_val calculated through conversion of Chi2 values to F values (Chi2/x lags) and conversion of F values through F.VERT.RE(F value, x lags, sample size) into p values
Coulibaly 2012	- Assumption: Wald test is expressed through Chi2 effect_typ
	- NCDEX: National Commodity Exchange in India
Chakraborty	- Calculated sample size on daily base
and Das 2013	- For GC test data on page 15 identification of lags was not possible. Df 1 and 2 are indicated but not further specified. I assumed df1 (10) consists of 3 lags for SV, TV, OI and 1df for the constant and I assume all 3 lags

Study	Coding protocol
	are equal, therefor every lag is assumed to be 3. DF2 might in indicate different sample sizes, but no further information are available and no further conclusions can be drawn, so I worked with the calculated sample size.
Ciner 2002	- Modified Baek & Brock gc test is a Hiemstra & Jones 1994 test - Calculated sample size on daily base - GC data on page 210 is "linear Granger causality test" data -> Assumption: Standard GC - Assumption: Lag length indicated is valid for m and n - Assumption: Hypothese_style is single, because coefficients are "jointly equal zero" - For non-linear Granger causality test values (BB GC) from asymptotically distributed N(0,1), p-values calculated
Coleman and Dark 2012	- Strange indication of values - Assumption start date is 01.01.1995 and end date is 01.01.2010 because dates are only on an annual basis - Calculated sample size on monthly base - Assumption: A Wald test is used so a Chi2 effect typ can be assumed - Assumption: Hypothese_style is single - Assumption: VECM lag order is assumed to be valid for GC test too. Lag length indicated is valid for m and n
Ding et al. 2014	- Assumption exact dates for sub period 1996 – 2003 is 01.01.1996 and 31.12.2003 because dates are only specified as first week of 1996 and last week of 2003 - Assumption exact dates for sub period 2004 – 2012 is 01.01.2004 and 31.12.2012 because dates are only specified as first week of 2004 and last week of 2012 - Calculated sample size on weekly base - Assumption: Wald test is used so a Chi2 effect typ can be assumed - Assumption: Lag length indicated is valid for m and n
Ederer et al. 2013	- Difficult to decide on period. In the paper they write they conducted the analysis on weekly base, but the underlying price date for coffee are only available on monthly base. Assumption: Weekly is correct. - Calculated sample size on weekly base - Assumption: Wald test is used so a Chi2 effect typ can be assumed - Assumption: Lag length indicated is valid for m and n - Assumption: Hypothese_style is single
Etienne et al. 2017	- Multivariat GC test included, but Hypothesis test for Bubble behavior or including Bubble behavior (page 60, table 6, page 61 table 7 ff) - Multivariat GC test uses SUR as "system" parameter - Assumption: All standard GC tests did not specify the "system" parameter, hence autoregressive distributed lag (ADL) is assumed - Calculated sample size on weekly base - They modified the GC approach to account for special bubble behavior
Fagan and Gencay 2008	- Data from text - Assumption: Wald test is used so a Chi2 effect typ can be assumed - Assumption: Hypothese_style is single

Fujihara and Mougou 1997	- Modified Baek & Brock GC test is a Hiemstra & Jones 1994 test
	- Lag length assumed, due to indication in index of phi for t-tests
	  o Data_id=1; m=9; n=31
	  o Data_id=2; m=1; n=40
	- Assumption: Lag length indicated for nonlinear GC is valid for m an n
	- Assumption: Sample size for standard GC is also valid for nonlinear GC (same period)
	- P-values calculated
	- Assumption: Hypothese_style is single

Study		Coding protocol
	-	Assumption: Exact dates for period January beginning 2000 to last week of 2010 are assumed to be 03.01.2000 and 30.12.2010
	-	Assumption: Lag length indicated is valid for m and n
Ghalayini 2011	-	Assumption: Hypothese_style is single
	-	Assumption that lag length indicated is valid for m and n
	-	D(logP), D(logF1), D(logF2), D(logF3), D(logF4) and D(logINV) not clearly specified. I assume D(logF1) to D(logF4) is for first differences of future price series for contract for 1 month till contract for 4 months, D(logP) is for first differences of spot price series and D(logINV) is for first differences of inventories
	-	Is not a total replica of Ghalayini 2011, but an extension
	-	Assumption: Exact dates for period January beginning 2000 to last week of 2010 are assumed to be 03.01.2000 and 30.09.2014
	-	Assumption: Lag length indicated is valid for m and n
Ghalayini 2017	-	Assumption: Hypothese_style is single
	-	Replication of calculations of Sanders and Irwin 2011, but with different time frame (3 days difference) still it is assumed to be a replication
	-	Calculated sample size on weekly base
	-	Assumption, that “CIT positions” are equal to “index returns” and both terms are used for each other
	-	Assumption, on page 10 table 5 the lower part data are F-statistics, because no OLS SE or robust SE is indicated
	-	Annotation: OLS SE and robust SE are available
	-	Calculated sample size on weekly base
	-	Assumption: Standard GC
Gilbert and	-	Assumption: Joint indicates soybean and soybean oil are calculated together
Pfuderer 2012	-	Assumption: Lag length indicated is valid for m and n
	-	Price returns are interpreted as logarithmic returns on nearby future
	-	Index returns are interpreted as absolute net long CIT positions
	-	Assumption: Sample sizes are indicated in the index of t-statistics and for F-statistics the m lag and sample size are indicated in the index.
	-	Start date for data on page 316 is assumed to be 03.04.2006 due to indication of April 2006
	-	Calculated sample size on weekly base
	-	Assumption: Lag length indicated is valid for m and n
Gilbert and Pfuderer 2014	-	Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	-	Assumption: Standard GC
Gilbert 2009	-	Assumption: VAR or ADL possible, but I assume VAR is used
	-	Assumption, on page 41 they say the Granger causality data are in table 7. Since there is no table 7 in the paper I assume they meant table 6.
	-	Assumption, on page 41 they say the first line of table 7 (here table 6) is H0 – 1, but in table 6 there are two H0 - 2 . I assume they meant row one of table 6 to be H0 – 1 cause two times H0 – 2 with different values makes no sense.
	-	Assumption: Tail probability is equal to p value
	-	Assumption: Standard GC
	-	Assumption: Lag length indicated is valid for m and n
Gilbert 2010a	-	Assumption: Hypothese_style is single for H0-1 & H0-2 and sum for H0-3 & H0-4
	-	Calculated sample size on quarterly base
	-	Assumption: Tail probability is equal to p value
	-	Assumption: Standard GC
Gilbert 2010b	-	Assumption: Lag length indicated is valid for m and n

Study	Coding protocol
	- Not clear if data on page 486 table 1 are all Granger test (like the heading suggests). Here it is assumed all of the table data are Granger test results
	- Assumption: Standard GC
	- Assumption: Lag length indicated is valid for m and n
Girardi 2015	- Assumption: Due to "all variables taken in daily percent changes" dif is assumed for lev_dif
	- Assumption table 5 panel A: Granger causality test; H0: Future unexpected trading volume does not Granger causes spot volatility and panel B vice versa, spot volatility does not Granger causes future unexpected trading volume
	- Assumption table 6 panel A: Granger causality test; H0: Futures unexpected open interest does not Granger cause spot volatility and panel B vice versa, Spot volatility does not Granger cause futures unexpected open interest
	- Assumption: Lag length indicated is valid for m and n
	- Calculated sample size on daily base
	- Assumption: Hypothese_style is single
Gupta et al. 2018	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- Assumption: Hypothese_style is single
Haase et al. 2018	- Assumption: F-test was used (t-test for underlying VAR) - Assumption: Lag length indicated is valid for m and n
	- Assumption: Standard GC
	- Assumption: Hypothese_style is single
	- Assumption: Lags chosen by ng-Perron criterion for ADF test are valid also for the GC test
Hannesson 2012	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Page 26, 09s1 is assumed to be 12.09.2009
	- Assumption: linear GC is standard GC
	- Calculated sample size on weekly base
	- Assumption: Lag length indicated is valid for m and n, excepted for
	- Assumption: Hypothese_style is single
	- "System" parameter for DP GC is VAR
Hernandez and Torero 2010	- Assumption: For all standard GC tests no "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed
	- Assumption: Standard GC
	- Assumption: Exact dates for period April 1998 to December 2013 are assumed to be 01.04.1998 and 31.12.2008
	- Calculated sample size on weekly base
	- Assumption: Hypothese_style is single
	- No specific lags were identified
Huchet and Fam 2016	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- Calculated sample size on weekly base
	- Assumption: Wald test indicates Chi2
	- No specific lags were identified for commodity_name "system"
Irwin and Sanders 2010	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC

Study	Coding protocol
	- Assumption: Standard GC
	- "lagged growth In ETF contracts", "lagged growth in ETF notional value", "contemporaneous growth in ETF contracts" is assumed to be a speculation position because ETFs are normally index replicating products not used by commercials for hedging.
Irwin and Sanders 2012	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Exact dates for period 1995 to 2006 are assumed to be 01.01.1995 and 31.12.2006
	- Assumption: Time series is assumed to be dif because "changes of positions" are indicated
	- Calculated sample size on weekly base
	- Assumption: Hypothese_style is single
	- No specific lags were identified
Irwin et al. 2009	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Carry effect is assumed to be the z variable
	- Hypothesis test on "only carry effect" not coded
	- Calculated sample size on weekly base
	- Assumption: Hypothese_style is single
Irwin et al. 2011	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Data not coded, because here a ECM instead of a classical Granger causality test is applied.
	- Error-Correction Model (ECM) for Testing for Long-Term and Short-Term Relationship for Price and Volume of Agricultural Futures Contracts in table 5 is assumed to be a Granger causality, it is subsumed in the later analysis under "linear GC"
	- Assumption: VAR is assumed
	- Assumption: Lag length indicated is valid for m and n, lag of three is indicated on page 59
	- Assumption: Hypothese_style is single
Malliaris and Urrutia 1998	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Data coded, but period not clear, start and end date for data not clear
	- Not clear if GC data useful, because no commodity in detail is examined but a commodity fund (GSCI)
	- Assumption: First differences are assumed also for GC tested data, because first differences was used in VAR for stationarity tests. The same is assumed for log func_form.
	- Assumption: Standard GC
Mathur et al. 2013	- Assumption: No date data indicated but assumed from grafics from 01.05.1990 till 01.05.2013 - Assumption: Hypothese_style is single
	- Assumption: Standard GC
	- Calculated sample size on monthly base
	- No specific lags for x variable were identified, but lags must be between 1 – 12 lags. (Lags for GARCH indicated but selected by out of-sample procedure not by AIC)
Mayer et al. 2017	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption exact dates for period June 2006 – June 2009 is 01.06.2006 and 30.06.2009 because dates are only specified as month and year
	- Calculated sample size on weekly base for palladium, because no data is available from August 2000 to September 2002.
Mayer 2012	- Assumption: Lag length indicated is valid for m and n

Study	Coding protocol
	- Calculated sample size on monthly base
	- Assumption: Hypothese_style is single
	- Assumption: Lag length indicated is valid for m and n (lag of 1 for x variable indicated in formula xt-1)
Merino and Albacete 2010	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- Assumption: First differences are not only calculated for returns but also for volatility
	- Assumption: Volatility calculation is also logarithmic
	- Assumption: Time period of "2004 to August 2010" is assumed to be 01.01.2004 to 31.08.2010
	- Calculated sample size on daily base
Mukherjee 2011	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Hypothese_style is single
	- Assumption that lag length indicated is valid for m and n
	- Assumption: Wald test indicates Chi2
Naderian and Javan 2017	- Assumption: Page 38 indicates lag length of 1 for m and n for nonlinear Granger causality test results
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Standard GC
	- Assumption: Time period of "2015 to June 2017" is assumed to be 02.01.2015 to 30.06.2017
	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Hypothese_style is single
Obadi and Korcek 2018	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Periodicity is strange (tue-tue, etc.) Check again!
	- Assumption: Hypothese_style is single
Often and Wisen 2013	- No specific lags were identified
	- Period might be different due to strange indication
	- Date data indicated in table 3 are not clear. Assumption:
	- Start / end dates are indicated from figure 1 as 02.02.2004 and 02.07.2010
	- Assumption of soy subperiodes: o 02.02.2004 – 02.02.2007 o 03.02.2007 - 15.08.2008 o 16.08.2008 – 02.07.2010
	- Assumption of corn subperiodes: o 02.02.2004 – 14.01.2005 o 15.01.2005 – 15.12.2006 o 16.12.2006 – 10.10.2008 o 11.10.2008 – 02.07.2010
	- Assumption: Data are first differences, because pretest data is first differenced
	- Next to k (assumed to be m lags) d is indicated but not explained, it is assumed d is n lags
	- Toda and Yamamoto (1995) procedure following the Rambaldi and Doran (1996) approach
	- No log or ln indicated there for linear func_form is assumed
Peri et al. 2013	- Calculated sample size on weekly base
	- Calculated sample size on weekly base
	- Assumption: Lag length indicated is valid for m and n
Prokopczuk et al. 2014	- Assumption: Some p-values are 0. Those p-value are calculated from the F values through F.VERT.RE(F value, x lags, sample size) into p values

Study	Coding protocol
	- Assumption: Standard GC
	- Calculated sample size on weekly base
	- "long-term GC test" p. 529 table 3 indicates a "t-statistic" but under equation 1 it is stated that hypothesis tests are done using F-tests. Therefor F-statistics are assumed for the p-values.
	- Autoregression is not explicitly coded, data_id 17 – 24 indicate hypothesis test testing for return and position
Sanders and Irwin 2011a	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- 0.1694 appears several times, which is strange, might be a calculation error of the author, is omitted
	- Chi2 assumed due to Wald test
	- "System" identifies the combined results for corn, soybean, soybean oil, cbot wheat, kcbot wheat, cotton, live cattle, feeder cattle, lean hog, coffee, sugar, cocoa, crude oil, natural gas
Sanders and Irwin 2011b	- No specific lags were identified for commodity_name "system"
	- Calculated sample size on daily base
	- Assumption: Standard GC
	- Page 25 suggests t-statistic, but F-statistic is identified before, hence F-statistic is assumed
Sanders and Irwin 2014	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Calculated Sample size on weekly base
	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
Sanders et al. 2004	- Assumption: Standard GC
	- Assumption: Exact dates for period 1995 to 2006 are assumed to be 03.01.1995 and 29.12.2006
	- Assumption: Standard GC
	- Long-term GC test (following Jegadeesh 1991) is included
Sanders et al. 2009	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Data for rolling moving window GC is ignored, because no values are indicated
	- Assumption: Hypothese_style is single
Sassi and Werner 2013	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Date data is not uniformly indicated
	- Assumption that lag length indicated is valid for m and n
	- Assumption: Standard GC
	- Calculated sample size on daily base
	- Assumption: Hypothese_style is single
	- No specific lags were identified
Sehgal et al. 2012	- Assumption: No "system" parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- Assumption that lag length indicated is valid for m and n
	- Unsure of date indication, period indication etc.
	- Assumption: Lag length indicated is valid for m and n
Sen and Paul 2010	- Assumption: Hypothese_style is single
	- Assumption: For data point 1 - 4 no start or end date are indicated, therefor the dates of Appendix 2 period 1 01.01.2003 till 29.05.2009 are assumed to be valid.

Study	Coding protocol
	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Hypothese_style is single
Serrao 2016	- No specific lags were identified
	- Problematic and partly not logical results (for example sample size 7 times under calculated sample size)
	- Assumption: Standard GC
	- Assumption: “Sum of coefficients of lagged values of the independent variable” is equal to p value
	- Assumption: Lag length indicated is valid for m and n (“Optimal number of lags L”)
	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed,
Shanker 2017	because it is a standard GC
	- Assumption: Exact dates for period January 2006 to September 2011 are assumed to be 02.01.2006 and 30.09.2010
	- Assumption: Data is in a weekly periodicity, because underlying data are from de CFTC DCOT and those data are on a weekly basis
Shanmugam	- Assumption: One asterisk indicates a rejection at 10% level of significance
and Armah	- Assumption: Hypothese_style is single
2012	- No specific lags were identified
	- Assumption: Exact dates for period 2004 to 2011 are assumed to be 02.01.2004 and 30.12.2011
	- Calculated sample size on daily base
Sharma and	- Assumption: Hypothese_style is single
Malhotra 2015	- Assumption: Lag length indicated is valid for m and n
	- Assumption: Exact dates for period 2004 to 2013 are assumed to be 02.01.2004 and 31.12.2013
	- Assumption: Atandard GC
	- Assumption: Hypothese_style is single
	- No specific lags were identified
	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed,
Sharma 2016	because it is a standard GC
	- Assumption: Exact dates for period January 2006 to July 2009 are assumed to be 03.01.2006 and 30.01.2009
	- Assumption: Cotton is listed two times, but ticker symbol KC is not cotton but coffee, so coffee instead of 2nd cotton is assumed
	- 4 influencing variables are indicated in GC test but it is assumed that only 2 are valid and the other 2 are indicating lags, so no z variable is assumed
	- Assumption: Lag length indicated is valid for m and n, so m and n are assumed to be 2
	- Assumption: Hypothese_style is single
Stoll and	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed,
Whaley 2010	because it is a standard GC
	- Assumption: Standard GC
	- Assumption: Exact dates for period January 2006 to December 2010 are assumed to be 03.01.2006 and 30.12.2010
	- Calculated sample size on daily base, 60 minute base, 30 minute base and 5 minute base
	- Assumption: Lag length indicated is valid for m and n, where date was available (AIC test data not available)
	- Assumption: Zero coefficient restriction test is assumed to be a Chi2 test
	- Table 7 might also be Granger causality test for impact through all other commodity markets (not just market j, Eq 1), like indicated in Eq 2, not coded, because not clear if GC or other procedure
Tse and	
Williams 2013	- P. 371 indicates 5 lags for day, 12 for 5-minute, 2 for 30-min and 1 for 60-min. But several lags where chosen by the AIC and are not included.

Study	Coding protocol
	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC
	- Assumption: Standard GC
	- Calculated sample size on daily base
	- Assumption: Lag length indicated is valid for m and n
Yang et al. 2005	- Assumption: No “system” parameters were indicated, hence autoregressive distributed lag (ADL) is assumed, because it is a standard GC

Appendix F: Data and R Code

The complete data set and the R code used to calculate the results of this master thesis are available under the following links:

Megastore University of Augsburg:

Excel table with coded data:

https://megastore.uni-augsburg.de/get/uaCB1_nytx/

Excel table with literature information:

https://megastore.uni-augsburg.de/get/CWz7eu9EwV/

R Code:

https://megastore.uni-augsburg.de/get/eByGj4LrWk/

Package list:

https://megastore.uni-augsburg.de/get/EOzwNafA92/

Github:

https://github.com/BongoKing/mra-granger

Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat. Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Martin Udo Hillenbrand

Augsburg, 30.04.2020