BANCO CENTRAL DE RESERVA DEL PERÚ

Unit roots in real primary commodity prices? A meta-analysis of the Grilli and Yang data set

Diego Winkelried*

* Universidad del Pacífico

DT. N°. 2017-013
Serie de Documentos de Trabajo
Working Paper series
Diciembre 2017

Unit roots in real primary commodity prices? A meta-analysis of the Grilli and Yang data set*

Diego Winkelried

Universidad del Pacífico (Lima, Perú)
winkelried_dm@up.edu.pe

This version: December 26, 2017

Abstract

The long-run behavior of real primary commodity prices, especially whether these series are trend stationary or contain a unit root, has been a topic of major debate in applied economics. In this paper, we perform a meta-analysis and combine the evidence of twelve representative studies on the subject, published in the last 25 years, in order to reach a unified conclusion about the presence of unit roots in these prices. The studies use different testing procedures, but share the common null hypothesis of a difference stationary process. Also, they use the individual price indices from the Grilli and Yang data set, arguably one the most popular source of long-term commodity price data. The combined evidence against unit roots in real primary commodity prices is strong: out of 24 cases, the unit root cannot be rejected in at most four. This implies that real primary commodity prices tend to be mean reverting and thus, to some degree, forecastable.

JEL Classification : C12, C22, O13.
Keywords : Primary commodity prices, unit roots, Grilli and Yang data, meta-analysis.

*I am indebted to Marco Terrones and seminar participants at the Universidad del Pacífico, the 2017 Annual Congress of the Peruvian Economic Association, and the XXXV Economists Meetings of the Central Reserve Bank of Peru, for their useful comments and suggestions. The financial support of the Research Center of Universidad del Pacífico (CIUP) is gratefully acknowledged. I alone am responsible for the views expressed in this paper and for any remaining errors.

1 Introduction

The long-run behavior of real primary commodity prices is a subject of major interest in development and international economics. Understanding the dynamic properties of these prices can have important policy implications for commodity-exporting countries, as the adequacy of export diversification or industrialization efforts, as well as the design of effective countercyclical macroeconomic policies, depend greatly on whether their terms of trade are expected to increase or decline sustainably in the future. For brevity, we will simply use "commodity prices" to refer to the real or relative price of primary commodities.

Since the early contribution of Cuddington and Urzua (1989), a growing body of literature has investigated whether commodity prices are trend stationary (TS henceforth) or if they are better characterized by difference stationary (DS henceforth), or unit root, processes (see Winkelried, 2018, for a recent review). This work has paralleled the methodological advances in the unit root testing literature (see Choi, 2015), which often find commodity prices to be a fertile ground for the empirical application of the testing procedures, and has prominently centered on the analysis of the price indices from the famous Grilli and Yang (1988) data set (GY, henceforth). The evidence against unit roots in commodity prices is, in general, mixed and broadly spread among individual studies. However, the existence of several studies tackling the same core question with similar data permits us to formalize a cross-paper comparison and to combine this dispersed evidence through a meta-analysis. This paper offers such meta-study, to provide a unified conclusion regarding the presence of unit roots in commodity prices. To this end, we gather the results from twelve representative studies published in the last 25 years, all of them using updated versions of the GY data set. Although their empirical explorations are based on different testing strategies, they share the null hypothesis of DS commodity prices.

There are two reasons why testing for unit roots in commodity prices is of interest. First, in development economics, there has been a widespread debate around the so-called Prebisch-Singer hypothesis, independently proposed by Prebisch (1950) and Singer (1950), that the relative prices of primary commodities in terms of manufactures are driven by a secular downward trend. In fact, the work by Grilli and Yang (1988) notably contributed to this discussion by bridging the information gaps that hindered the empirical assessment of trends in commodity prices (see Ghoshray, 2011, for a historical account). It is also widely recognized that ignoring a unit root can have profound distortionary effects on the inferences made about the long-run behavior of a time series. In particular, the so-called "spurious trend" problem (i.e., when fitting a DS series to a linear trend, the slope coefficient would appear highly significant, even though the series may not contain a trend at all) may emerge. Thus, careful modeling of the statistical processes underlying commodity prices seems essential for the correct empirical assessment of key conjectures such as the Prebisch-Singer hypothesis (see Cuddington et al., 2008, for a comprehensive survey).

Second, in a sequence of influential studies, Deaton and Laroque (1992, 1996, 2003) and Deaton (1999) argue that there are no compelling reasons for commodity prices to exhibit a trending behavior. Moreover, their theoretical findings, derived from models with competitive and speculative storage, or from models where supply is infinitely elastic due to the abundance of labor at a subsistence wage in commodity exporters, also predict that commodity prices should be stationary or mean-reverting. This implies that shocks to real commodity prices, that may be persistent, are transitory by nature. Despite being theoretically sound, such predictions would be at odds with the empirical evidence if we are unable to reject the null hypothesis of nonstationary commodity prices.

Deaton and Laroque (2003) argue that in practice the stationarity of commodity prices may receive limited empirical support because unit root tests often lack statistical power in finite samples: the (false) null hypothesis of a unit root may not be rejected against stationary but "persistent" alternatives (see Choi, 2015, ch. 3). In addition, when the alternative model ignores certain features, other than the unit root, that may affect the long-run behavior of a time series, a DS model can result in a better characterization of the data. Leading examples are unmodeled structural breaks or some forms of neglected nonlinearities. In other words, such tests may fail to reject the unit root hypothesis not because of the merits of the null hypothesis, but because of the inadequacy of the alternative hypotheses.

The lack of power of unit root tests is perhaps the most important reason why a meta-analysis on the stationarity properties of a group of time series is of value. As stressed by Stanley (2001), beyond providing a formal method to conduct systematic reviews and comparisons, the main advantage of a meta-analysis is the increase in power that follows from combining dispersed, and possibly conflicting, evidence. If the n studies involved turn out to be independent, statistical power would increase corresponding to the sum of the sample sizes across all n studies. The increase in power is expected to be less striking in our analysis, as the collected studies are hardly independent, but the combining evidence still delivers significant gains when compared to the individual studies (see Demetrescu et al., 2006).

Our analysis focuses on the the GY data set due to its popularity and its great influence on the academic debate about the long-term behavior of commodity prices. It features annual data from 1990 for 24 primary commodity nominal prices (11 foodstuffs, 7 nonfood soft commodities and 6 metals) in US dollars, and a manufacturing unit value index, also in US dollars, that serves as a deflator: the series of interest are the logarithms of the ratios of the commodity prices to this deflator. The data also include aggregate indices.[1] However, as argued in Cuddington (1992), not only the weights using in the aggregation are somehow arbitrary and not necessarily representative of the terms of trade of primary exporters, but also the aggregate indices often display rare dynamic features (such as structural breaks) that are difficult to reconcile with those of the individual prices. Thus, following the tradition of most empirical studies, our interest is on each individual commodity price.

Of course, not all studies on the dynamic properties of commodity prices, such as their persistence, use the GY data. Some examples are Cashin et al. (2000) who use IMF data, Enders and Holt (2012) who use the World Bank's "pink sheet", and Harvey et al. (2010) who construct an entirely new long-term data set from a number of sources. We exclude these studies from the meta-analysis to enhance the cross-study comparability, as the prices included, the primary sources of information, the deflators, the frequency of the observations, among others, differ from one study to the other. On the other hand, not all studies using the GY data necessarily perform unit root tests, for instance Winkelried (2016) and Gouel and Legrand (2017), and naturally cannot be included in the meta-analysis.

Our methods are somewhat related to panel data procedures, such as those in Iregui and Otero (2013) and Arezki et al. (2014). A key difference is that panel methods are implicitly based on a factor structure that is shared by all commodities, so cross-sectional dependence may arise across commodities. The meta-analysis exploits a different source of variation, the heterogeneity across studies, and thus dependence may arise across studies for a given commodity price, *regardless* on how it relates to another commodity price. However, panel methods can also enhance the power of individual unit root tests (see, *inter alia*, Choi, 2001) and the few applications on real primary commodity prices strongly reject nonstationarity. This, in fact, is also the main conclusion of our study.[2]

The rest of the paper is organized as follows. Section 2 presents the empirical studies included in our meta-analysis, along with a brief review of the 14 unit root tests used in these studies. Section 3 discusses how to combine the evidence of several studies, especially when the results among studies are expected to be statistically dependent. This section also describes the data used in the meta-analysis, which is a collection of p-values from the various unit root tests, and presents the main results. The evidence against unit root in commodity prices is strong. Out of 24 commodities, the DS hypothesis is categorically rejected in 20 cases, and cannot be strongly rejected in only one case. In the remaining three cases, the conclusion depends on the significance test used and on whether outliers are included or excluded from the computations. Section 4 concludes and provides suggestions for future research.

[1] The aggregate indices have been used in studies such as Cuddington and Urzua (1989), Ardeni and Wright (1992), Bleaney and Greenaway (1993), Zanias (2005), Cuddington et al. (2008) and Mariscal and Powell (2014).

[2] Iregui and Otero (2013) also use the GY data. However, they only report aggregate, panel-wide statistics so it cannot be included in the meta-analysis.

2 Review of the literature

In this section we provide a review of the literature. First, we describe the statistical procedures used in the studies involved in our meta-analysis. Specifically, these are 14 different unit root tests. Then, we present the 12 studies using the tests to assess the nonstationarity of commodity prices.

2.1 Unit root tests

Next we present a brief account of the unit root tests used in the empirical studies under consideration. Choi (2015) provides a comprehensive textbook treatment on these methods, including topics from the abundant unit root testings literature that are not covered here.

In what follows, y_t denotes the logarithm of the relative commodity price of interest. The tests are often derived from the output of linear regressions with T observations, and we use e_t to denote a generic error term in such regressions. Also, the estimated equations usually include a number of deterministic functions of t, collected in vector $\boldsymbol{w}_t$. Among these are dummy variables that control for the presence of structural breaks. We denote the location of a break by $\lambda \in (0,1)$ such that $d_t(\lambda) = 1$ if $t > T\lambda$ and $d_t(\lambda) = 0$ if $t \leq T\lambda$ is the dummy variable associated with a level shift, and $D_t(\lambda) = (t - T\lambda)d_t(\lambda)$ is the dummy variable associated with a change in slope.

2.1.1 OLS tests

The standard unit root tests are obtained from the output of the OLS estimation of the equation

$$\Delta y_t = \boldsymbol{\psi}' \boldsymbol{w}_t + \phi y_{t-1} + e_t \,, \tag{1}$$

The Dickey and Fuller (1979) test that includes an intercept and a linear trend uses $\boldsymbol{w}_t = (1, t)'$. The unit root statistic is the t-statistic for the significance of ϕ in (1), that we denote as τ_ϕ. The null hypothesis of a DS process is rejected, in favor of a TS around a linear time trend, for large negative values of τ_ϕ.

It is well-known that the distribution of τ_ϕ under the null hypothesis does not depend on nuisance parameters when the error term in (1) is serially uncorrelated. Said and Dickey (1984) establish the widespread practice, adopted in many of the subsequent unit root tests, to augment equation (1) with a suitable selection of lags of Δy_t in order to ensure that this property is satisfied, rendering the Augmented Dickey and Fuller test (ADF henceforth). Phillips and Perron (1988) propose a corrected t-statistic instead. In either case, the asymptotic distribution of τ_ϕ under the null hypothesis would then be identical to that of the regression with uncorrelated errors, thus free of nuisance parameters.

On the other hand, the interest on how the presence of structural breaks affect the properties of unit root tests begins with Perron (1989). This author shows, in particular, that the power of such tests can be severely reduced: a TS series around a broken trend may appear as a DS process if the structural change is ignored during testing. Thus, assuming that λ is known, Perron (1989, PER henceforth) suggests to use $\boldsymbol{w}_t(\lambda) = (1, t, d_t(\lambda), D_t(\lambda))'$ in (1) and to base the inferences on the t-statistic $\tau_\phi(\lambda)$. Unlike the ADF test, PER allows structural changes also under H_0; i.e., a DS process with a changing drift against a TS process around a broken linear trend.

A critique to PER, however, is that the limiting distribution of $\tau_\phi(\lambda)$ depends on λ, a nuisance parameter that may be unknown in practice. Zivot and Andrews (1992, ZA hereafter) deal with the problem of testing the null of a DS process against a TS process under a structural break of unknown date, by performing a grid search over possible values of λ and employing $\tau_{\text{ZA}} = \inf_{\lambda \in (0,1)}\{\tau_\phi(\lambda)\}$ as the relevant test statistic. The limiting distribution of τ_{ZA} does not depend on λ, and structural breaks only feature under H_1.

Lumsdaine and Papell (1997, henceforth LP) extend the ZA procedure and allow for two structural breaks. In particular, for $\lambda_2 > \lambda_1$, they use $\boldsymbol{w}_t(\lambda_1, \lambda_2) = (1, t, d_t(\lambda_1), D_t(\lambda_1), d_t(\lambda_2), D_t(\lambda_2))'$ in (1) and consider the minimum t-statistic of ϕ over a two-dimensional grid, i.e., $\tau_{\text{LP}} = \inf_{\lambda_1 \in (0,1), \lambda_2 \in (\lambda_1, 1)}\{\tau_\phi(\lambda_1, \lambda_2)\}$. As in ZA, the structural breaks appear only under H_1 and the limiting distribution of τ_{LP} does not depend on nuisance

parameters either.

Finally, Enders and Lee (2012, EL hereafter) propose a test meant to be used when the dates, form and number of structural breaks are unknown. Instead of introducing dummy variables to account for such changes, they consider to model the deterministic term using a Fourier expansion. Even though such expansion is particularly suitable to approximate "smooth breaks", these authors argue that it does a decent job also in approximating "sharp breaks". In particular, their basic specification has

$$\boldsymbol{w}_t(k_1, k_2) = \left(1, t, \sin\left(\frac{2\pi k_1 t}{T}\right), \cos\left(\frac{2\pi k_1 t}{T}\right), \sin\left(\frac{2\pi k_2 t}{T}\right), \cos\left(\frac{2\pi k_2 t}{T}\right)\right)',$$

where k_1 and k_2 represent the frequencies used to filter out the effects of structural breaks and other nonlinearities that may be mistaken for a unit root. Often $k_1 = 1$ and $k_2 = 2$. The limiting distribution of $\tau_\phi(k_1, k_2)$, which is free from nuisance parameters, depends on the chosen frequencies in the Fourier terms, i.e., the values of k_1 and k_2.

2.1.2 LM tests

Schmidt and Phillips (1992) note that the finite sample power of tests such as ADF could be enhanced if the coefficients of the deterministic component $\boldsymbol{\psi}$ were estimated more efficiently. In the ADF test $\boldsymbol{\psi}$ is implicitly estimated from a regression of y_t on $\boldsymbol{w}_t$; however, under $H_0 : \phi = 0$ it would be more appropriate to consider $\hat{\boldsymbol{\psi}}$, the estimator of $\boldsymbol{\psi}$ from a regression of Δy_t on $\Delta \boldsymbol{w}_t$ (i.e., a restricted estimator for $\phi = 0$). Define $\tilde{y}_t = y_t - \hat{\boldsymbol{\psi}}'\boldsymbol{w}_t - (y_1 - \hat{\boldsymbol{\psi}}'\boldsymbol{w}_1)$ as a detrended version of y_t, with $\tilde{y}_1 = 0$. Then, the unit root statistic, which we call τ_{LM}, is the t-statistic for the significance of ϕ in

$$\Delta y_t = \boldsymbol{\psi}'\Delta \boldsymbol{w}_t + \phi \tilde{y}_{t-1} + e_t\,, \tag{2}$$

and is proportional to the score vector evaluated at H_0, hence the name of the procedure.

Lee and Strazicich (2003, LS hereafter) extend the LP two-break test within this LM framework. As in LP, they use $\boldsymbol{w}_t(\lambda_1, \lambda_2) = (1, t, d_t(\lambda_1), D_t(\lambda_1), d_t(\lambda_2), D_t(\lambda_2))'$ and consider $\tau_{\text{LS}} = \inf_{\lambda_1 \in (0,1), \lambda_2 \in (\lambda_1, 1)}\{\tau_{\text{LM}}(\lambda_1, \lambda_2)\}$, where $\tau_{\text{LM}}(\lambda_1, \lambda_2)$ is the t-statistic of ϕ in (2). However, unlike LP, the null hypothesis in the LS test is a DS process subject to structural changes and, in general, the limiting distribution of τ_{LS} depends on the location parameters λ_1 and λ_2.

Finally, Meng et al. (2017, MLP from now on) refine the LS approach to deal with these nuisance parameters. In particular, they show that if $\tilde{y}_t$ in (2) is replaced by

$$\tilde{\tilde{y}}_t = \left(\frac{1 - d_t(\lambda_1)}{\lambda_1} + \frac{d_t(\lambda_1) - d_t(\lambda_2)}{\lambda_2 - \lambda_1} + \frac{d_t(\lambda_2)}{1 - \lambda_2}\right)\tilde{y}_t\,,$$

then the asymptotic distribution of τ depends only on the number of breaks, and no longer on their specific location. Concretely, they show that the distribution of the MLP test is the same as the distribution of $\tau_{\text{LM}}(0.33, 0.66)$ in the LS test.[3]

2.1.3 GLS tests

Elliot et al. (1996) popularize the so-called "GLS-detrending" method, also aimed to increase the finite sample power of unit root tests. They note that under H_1, the model in differences in the LM procedure is misspecified and propose to use quasi-differenced data instead: $y_t^a = y_t - a y_{t-1}$ and $\boldsymbol{w}_t^a = \boldsymbol{w}_t - a\boldsymbol{w}_{t-1}$ for $t = 2, \ldots, T$, and $y_1^a = y_1$ and $\boldsymbol{w}_1^a = \boldsymbol{w}_1$, where $a = 1 - c/T$, with c being a constant that is calibrated to improve the finite sample performance of the test. In the limit, $a \to 1$ as in LM-detrending.

[3] Meng et al. (2017) also consider augmenting $\boldsymbol{w}_t$ with functions of the residuals in (2) to allow for nonnormal errors, the so-called "residual augmented least squares" method. In their empirical application on the GY dataset, the results of the extended procedure are almost identical to the simpler test that we consider in the meta-analysis.

The detrended series is $\tilde{y}_t = y_t - \hat{\boldsymbol{\psi}}' \boldsymbol{w}_t$, where $\boldsymbol{\psi}$ is estimated from a regression of y_t^a on $\boldsymbol{w}_t^a$. This regression resembles a GLS correction for error autocorrelation, and hence the name of the procedure. The unit root statistic is the t-statistic for the significance of ϕ in

$$\Delta \tilde{y}_t = \phi \tilde{y}_{t-1} + e_t \,. \tag{3}$$

For a test with a linear trend, $\boldsymbol{w}_t = (1, t)'$, Elliot et al. (1996, ERS henceforth) suggest using $c = 13.5$, whereas Ayat and Burridge (2000, AB from now on) suggest using $c = 18.5$ when a quadratic trend is also included, $\boldsymbol{w}_t = (1, t, t^2)'$. Harvey et al. (2011) argue that allowing for a quadratic trend does not necessarily seek to obtain a realistic description of the data generating process, but appears to be a useful device to approximate nonlinearities in the unknown deterministic component of the series. Their motivation is similar to that of the EL test, with the important difference that the unknown deterministic components is approximated with further monomials of time rather than with Fourier terms.

Ng and Perron (2001, NP hereafter) propose a different type of test, which is effectively a modified Phillips and Perron (1988) test, also based on GLS-detrending. The test is no longer computed as the t-statistic of (3), but instead as

$$Z_\phi = \frac{(\tilde{y}_T)^2 - T s^2}{2\sqrt{s^2 \sum_{t=1}^{T} (\tilde{y}_{t-1})^2}}, \tag{4}$$

where s^2 is typically estimated as the variance of the error term in (3). As in ERS, they use $c = 13.5$.

Perron and Rodriguez (2003, PR henceforth) extend the NP test to the case of a single structural break. In the same spirit of ZA, they use $\boldsymbol{w}_t(\lambda) = (1, t, d_t(\lambda), D_t(\lambda))'$ and perform a grid search over λ to obtain $Z_{\text{PR}} = \inf_{\lambda \in (0,1)} \{Z_\phi(\lambda)\}$ as the test statistic, where $Z_\phi(\lambda)$ is computed as in (4) for a given value of λ. The limiting distribution of τ_{PR} does not depend on λ, but, unlike ZA, structural breaks are allowed under H_0. As in ERS, they use $c = 13.5$, and the limiting distribution of Z_ϕ is free from nuisance parameters.

Finally, Carrion-i Silvestre et al. (2009, CKP hereafter) extend the PR procedure to allow for two breaks, using $\boldsymbol{w}_t(\lambda_1, \lambda_2) = (1, t, d_t(\lambda_1), D_t(\lambda_1), d_t(\lambda_2), D_t(\lambda_2))'$ and $Z_\phi(\lambda_1, \lambda_2)$ as computed in (4). Structural breaks are allowed under both H_0 and H_1, and the limiting distribution of Z_ϕ depends on λ_1 and λ_2. The authors provide response surfaces to calibrate $c = c(\lambda_1, \lambda_2)$ for specific configurations of the breaks location.

2.1.4 Other procedures

Kapetanios et al. (2003, KSS henceforth) propose a unit root test that is powerful against a stationary but nonlinear alternative model. In particular, they consider the smooth transition autoregression $\Delta y_t = \phi y_{t-1} + \gamma y_{t-1} \Theta(\theta y_{t-d}) + e_t$, where $\Theta(.)$ is a nonlinear function of the lag y_{t-d} such that $\Theta(0) = 0$ and $\theta \geq 0$. The unit root statistic is the t-statistic of δ in the testing equation $\Delta y_t = \delta y_{t-1}^3 + \text{error}_t$ which is obtained as a Taylor approximation of the smooth transition autoregression around the null model.

Finally, Leybourne et al. (2007, LKT hereafter) test the unit root hypothesis against an alternative process with changing persistence; i.e., a nonlinear model that is subject to shifts between $I(0)$ and $I(1)$ regimes. Using GLS-detrended data for $c = 10$, they define $\tau(\lambda_1, \lambda_2)$ as the t-statistic on ϕ in equation (3) estimated with sample observations between $\lambda_1 T$ and $\lambda_2 T$, and consider $\tau_{\text{LKT}} = \inf_{\lambda_1 \in (0,1), \lambda_2 \in (\lambda_1, 1)} \{\tau(\lambda_1, \lambda_2)\}$ as the relevant test statistic. Its limiting distribution does not depend on λ_1 or λ_2.

2.2 Unit roots in the Grilli and Yang data set

We now present a quick review of the studies considered in our meta-analysis; in particular, the sample used, the testing strategy followed and the main conclusions regarding the nonstationarity of the commodity prices in the GY data set. This information is summarized in Table 1.

Following Cuddington and Urzua (1989), the seminal paper on unit roots in the disaggregated real price

indices of the GY data set is Cuddington (1992, Cud). He uses the very first release of the data, up to 1983, and based most of his analysis on the ADF test. This author also acknowledges the possible presence of structural breaks and consider the PER test, albeit only for the case of coffee. Out of the 24 commodities, the unit root is rejected for 12 (coffee, hides, lamb, lead, maize, palmoil, rice, sugar, timber, tin, wheat and zinc).

Using IMF data, León and Soto (1997, LeSo) extend the GY data up to 1992 and focus on the importance of adequately account for structural breaks when testing for unit roots. They apply the ZA test to the 12 commodity prices found to be DS by Cuddington (1992), and reject the unit root in 8 of these cases (aluminum, copper, cotton, jute, rubber, tea, tobacco and wool). For the remaining 12 commodities, found to be TS by the ADF test, León and Soto (1997) report results only for the 7 cases where the standard Chow breakpoint test rejected a linear trend with no breaks (coffee, maize, palmoil, rice, sugar, timber and tin). The results for the 5 commodities left (hides, lamb, lead, wheat and zinc) are taken from Kim et al. (2003) who, apart from their full-sample results, report the statistics for the subsample used by León and Soto (1997). This study is the first to present a strong case against unit roots by concluding that overall 20 commodity prices are better described by TS processes.

Kim et al. (2003) and Newbold et al. (2005) develop ARIMA modeling strategies to infer the presence of trends or drifts in commodity prices when there is uncertainty about the order of integration of the series. Using World Bank sources, they update the GY data, respectively, up to 1998 and 2002. When computing ADF tests in these updated samples, the evidence becomes less supportive for stationarity than in previous works. In both studies, the unit root is rejected only for 7 prices (aluminum, hides, lamb, rubber, sugar, timber and zinc), whereas in Kim et al. (2003) a weak rejection is also found for the price of lead.

Kellard and Wohar (2006) update the GY data up to 1998 (they do not report details on the sources) and extend the work of León and Soto (1997) by applying the LP test that allows, for the first time, for up to two structural breaks. The evidence against unit roots is strong but not overwhelming, as 14 series are found to be TS. In particular, the unit root is rejected against a linear trend with two breaks for 10 commodities (hides, lead, maize, rubber, silver, tea, timber, tin, wool and zinc), and against a linear trend with a single break in 4 cases (aluminum, jute, palmoil and rice). It is worth mentioning that Kellard and Wohar (2006) do not report the LP statistics for the 10 cases where the unit root was not rejected (banana, beef, coffee, cocoa, copper, cotton, lamb, sugar, tobacco and wheat). They do report, however, the results of the NP test, which are the ones we considered in the meta-analysis.

Pfaffenzeller et al. (2007) provide a detailed explanation of how to update the GY data using information from the World Bank up to 2003. More recent updates following this methodology are made publicly available at `http://www.stephan-pfaffenzeller.com/cpi.html`, and this is the source of information in all 7 subsequent studies in our meta-analysis.

Balagtas and Holt (2009, BH) uses data up to 2003, and is the first study to consider a nonlinear alternative model. In particular, they use the KSS approach to test for a unit root against a smooth transition autoregression. They present further support against nonstationarity as the unit root cannot be rejected in only 5 cases (banana, jute, maize, rice and wheat). It is worth stressing, however, that the nonlinear alternative models fitted to 4 prices (lamb, tea, tin and tobacco) display locally explosive behavior, raising doubts about the stationarity of these series.

Using also data up to 2003, Ghoshray (2011, Gho11) builds on the findings of Kellard and Wohar (2006) by using the more powerful one-break PR and two-break LS tests. The PR test rejects the unit root against a TS process with a single break in 16 cases. The LS test confirms this conclusion in 2 cases (banana and zinc), indicates that the rejection should be against a TS process with two breaks in 11 instances (copper, cotton, lead, palmoil, rice, rubber, timber, tin, tobacco, wheat and wool), and fails to reject the unit root for the remaining 3 commodities (jute, maize and sugar). The LS tests are only reported for the aforementioned 13 cases; for the 11 cases where the LS test does not reject, the PR results are considered in the meta-analysis.

Harvey et al. (2011, HLT) propose a union of rejections strategy which allows for a more flexible alternative model. In particular, the decision rule is to reject the unit root if either the ERS test against a linear trend or

the AB test against a quadratic trend rejects. Thus, in the meta-analysis the minimum between both statistics is considered. The inclusion of a quadratic trend provides a simple parametric way to approximate a linear trend that undergoes structural breaks at some unknown points. Using the GY data up to 2003, they find that 4 prices (beef, copper, lamb and sugar) can be taken as TS around a stable linear trend, whereas 12 commodities (aluminum, coffee, lead, maize, palmoil, rice, rubber, tobacco, timber, wheat, wool and zinc) are better characterized as TS around a nonlinear time trend.

Ghoshray (2013, Gho13) considers the nonlinear alternative, entertained in the KLT test, of a "dynamic persistence" process, i.e., a process shifting between $I(0)$ and $I(1)$ regimes. He argues that multiple changes in persistence are likely to be a better characterization of commodity price series due to a number of factors (political, technological, economic, among others) that may alter the nature of the data. Using the GY data up to 2008, only in 6 cases (coffee, cocoa, copper, lead, tea and silver) the unit root cannot be rejected.

Back to linear unit root testing under structural breaks, Ghoshray et al. (2014, GKW) deviate from the traditional approach of treating the determination of breaks and the presence of unit roots simultaneously, and proceed sequentially. They first test for up to two structural breaks using an inferential framework that is valid regardless on whether the possibly broken series is $I(0)$ or $I(1)$. Then, armed with the classification between stable and broken series, they perform, respectively, the NP and the CKP unit root tests. They find 16 commodities to be TS processes: 8 around a linear trend (beef, lamb, lead, rice, sugar, timber, wheat and zinc), 3 around a linear trend with a single slope break (coffee, cotton and jute), and 5 with two structural breaks (maize, palmoil, rubber, tea and tobacco).

Meng et al. (2017, MLP) is the least supportive study for unit roots in the meta-analysis. They apply their modification of the two-break LS statistic to the GY commodity prices up to 2007, and find that the unit root cannot be rejected in just 3 cases (palmoil, lamb and copper). They argue that, relative to other studies, this a manifestation of the virtues of the modified LS test: it is more powerful and its distribution is free from nuisance break-location parameters.

Finally, Winkelried (2018, Win) considers the EL test that allows for a flexible trend, approximated by Fourier terms, in the alternative model (the ADF test is encompassed by this approach). It is argued that the Fourier terms are able to control for the presence of hypothesized super cycles in the data, which are stationary but quite persistent, that otherwise would be mistaken by a unit root. Thus, like León and Soto (1997) and Meng et al. (2017), the evidence against unit roots is strong, as in only 4 cases (beef, lead, tin and silver) the DS model is not rejected.

3 Methodological issues and results

Next, we present a methodological discussion on how the evidence from different statistical tests are transformed into a common and comparable metric. Then, we describe the data collection and present descriptive statistics through a meta-regression analysis. Finally, we present the results on our meta-analysis on the presence of unit roots in commodity prices.

3.1 Combining evidence

The purpose of a meta-analysis is to combine the outcome of several studies to reach an overall conclusion about the significance of statistical tests of interest. Since the tests statistics can be arguably very different, sharing only an underlying null hypothesis, the units of comparison across studies are the corresponding p-values. Thus, the problem becomes a combination of significance levels. Even though there are a few ways to do so, we focus on the so-called inverse normal method that has gained prominence in panel data econometrics (see Choi, 2001; Demetrescu et al., 2006; Costantini and Lupi, 2013). As discussed in Cheng and Sheng (2017), this method performs well when the evidence against H_0 is broadly spread among the various tests. Its most important advantage, however, is that, as shown in Hartung (1999) and Demetrescu et al. (2006), the method can be extended in a relatively simple fashion to deal with dependence among the

underlying tests statistics.

The following exposition is for a given commodity price. Consider n studies indexed by $i = 1, 2, \ldots, n$, and let τ_i be the one-sided test statistic for testing the unit root hypothesis, where large negative values of τ_i lead to a rejection. If τ_i has a continuous distribution function $F_i(\cdot)$ under the unit root hypothesis, then the p-value $p_i = F_i(\tau_i)$ is uniformly distributed on $(0, 1)$ regardless of the form of $F_i(\cdot)$. Then, the probit or z-score $z_i = \Phi^{-1}(p_i)$, where Φ is the cumulative standard normal distribution function, satisfies $z_i \sim N(0, 1)$. Moreover, if the probits $\{z_1, z_2, \ldots, z_n\}$ are independent, then

$$z(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \sim N(0, 1),$$

and a combined p-value can be obtained from $p = \Phi(z(0))$. This approach is often known as Stouffer's method (see Cheng and Sheng, 2017). A major limitation is the assumption of independence, which may be inappropriate when the individual statistics are expected to be correlated because, among other reasons, they come from updated versions of the same data set. It is well-known that in this situation the combined test will have serious size distortions, rejecting the null hypothesis quite too often.

Hartung (1999) extends Stouffer's method to allow for dependency among test statistics. Given normality, dependency is equivalent to correlation and he considers the stylized case that $\mathrm{Cov}(z_i, z_j) = \mathrm{Corr}(z_i, z_j) = \rho$, for $i \neq j$ and for some real-valued parameter satisfying $-(n-1)^{-1} < \rho \leq 1$. Then, the combined z-score becomes[4]

$$z(\rho) = \frac{1}{\sqrt{1 + \rho(n-1)}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \right) = \frac{1}{\sqrt{1 + \rho(n-1)}} z(0), \tag{5}$$

which is also normally distributed under the null hypothesis. In the usual case that $z(0)$ is negative, $z(\rho)$ is increasing in the correlation ρ. In this setup, ρ can be interpreted as a measure of redundant information across studies,[5] so positive values of ρ imply that, relative to the independent case, the amount of new information brought by an additional study is limited; thus, $z(\rho)$ shrinks $z(0)$ towards zero accordingly, correcting the aforementioned size distortions.

The value of ρ is, however, unknown. A strategy to draw inferences in this situation is to determine an upper bound of ρ for rejection. More precisely, for a significance level α, define $\bar{\rho}_\alpha$ such that $z(\rho) = z_\alpha$, where z_α

[4] Hartung (1999) considers a more general statistic that attaches a weight $\omega_i > 0$ to the study i:

$$z(\rho) = \left[(1-\rho) \sum_{i=1}^{n} \omega_i^2 + \rho \left(\sum_{i=1}^{n} \omega_i \right)^2 \right]^{-1/2} \sum_{i=1}^{n} \omega_i z_i \,.$$

Usually the weights are proportional to the sample size used to compute z_i, following the premise that larger samples produce more precise estimates. It is unclear, however, whether weighting is required since the p-values underlying z_i, at least as computed here, already make allowances for different sample sizes. Despite this and the fact that we found very small effects of the sample size in the meta-regressions of Table 2 below, we computed weighted versions of the combined z statistic using both $\omega_i = T_i^{1/2}$ and $\omega_i = T_i$. No conclusion reached by using the unweighed statistic (5) was affected in any meaningful way, and thus we decided not to report these results, which are available upon request.

[5] To illustrate, consider a textbook example consisting of 3 studies each with T observations of a random variable x with $\mathrm{E}(x) = 0$ and $\mathrm{var}(x) = \sigma^2$. It is known that the sample is completely random in the first study, but in the second and third studies m observations are from the first sample while the remaining $T - m$ observations are drawn independently. Which observations are repeated, however, is unknown. Since the sample averages, $\bar{x}_i$, are unbiased estimators for the mean and $\mathrm{var}(\bar{x}_i) = \sigma^2/T$, the hypothesis $H_0 : \mathrm{E}(x) = 0$ can be tested with the statistics $z_i = \bar{x}_i / \mathrm{var}(\bar{x}_i)^{1/2} = T^{1/2} \bar{x}_i / \sigma$ for $i = 1, 2, 3$. Consider now the pooled estimator $\bar{x} = (\bar{x}_1 + \bar{x}_2 + \bar{x}_3)/3$ and note that $\mathrm{var}(\bar{x}) = \sigma^2(1 + 2m/T)/(3T)$, since $\mathrm{cov}(\bar{x}_i, \bar{x}_j) = \sigma^2 m/T^2$ for $i \neq j$. The pooled z-score for H_0 is then $\bar{z} = \bar{x}/\mathrm{var}(\bar{x})^{1/2} = 3^{-1/2}(z_1 + z_2 + z_3)/(1 + 2m/T)^{1/2}$, which has the form of $z(\rho)$ with $n = 3$ and $\rho = m/T$ (the proportion of repeated observations across samples; i.e., "redundant information"). If the studies have a different number of observations, the pooled statistic $\bar{z}$ has the form of the weighted statistic in footnote 4 with $\omega_i = T_i^{1/2}$ and $\rho = 3m/[(T_1 T_2)^{1/2} + (T_1 T_3)^{1/2} + (T_2 T_3)^{1/2}]$.

is the appropriate critical value from the standard normal distribution,

$$\bar{\rho}_\alpha = \frac{1}{n-1}\left[\left(\frac{z(0)}{z_\alpha}\right)^2 - 1\right].$$

Then, the null hypothesis would be rejected, $z(\rho) \leq z_\alpha$, for all ρ satisfying $\rho \leq \bar{\rho}_\alpha$. A large value of $\bar{\rho}_\alpha$, thus, constitutes strong evidence for rejection at the α significance level.

A more common approach is to estimate ρ from $\{z_1, z_2, \ldots, z_n\}$. Hartung (1999) suggests using $z(\hat{\rho})$, where

$$\tilde{\rho} = 1 - \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \frac{1}{n}\sum_{j=1}^{n} z_j\right)^2 \qquad \text{and} \qquad \hat{\rho} = \tilde{\rho} + \kappa\sqrt{\frac{2}{n+1}}(1-\tilde{\rho}).$$

Under Hartung's assumptions, $\tilde{\rho}$ is an unbiased estimator of ρ, and its sampling variance can be unbiasedly estimated by $\hat{\text{var}}(\tilde{\rho}) = 2(1-\tilde{\rho})^2/(n+1)$. Since ρ enters nonlinearly in $z(\rho)$, plugging $\tilde{\rho}$ into (5) turns out to underestimate the expectation of the denominator; hence, Hartung (1999) proposes to use $\hat{\rho}$ instead that corrects $\tilde{\rho}$ by adding a fraction of its standard deviation. The parameter $\kappa > 0$ regulates the actual significance level by avoiding the underestimation of the denominator. Among various alternatives, Hartung (1999) finds that the value $\kappa = 0.2$ works well in practice.

It is worth mentioned that Hartung's framework has been put to a test and generalized in Demetrescu et al. (2006). These authors show that the pairwise correlation of the individual z-scores need not be constant for the combined z-score to be valid. More importantly, they also note that the fact that the correlated probits are marginally normal by construction, does not imply that they are jointly multivariate normal, which is a relevant distributional assumption in Hartung (1999). Yet, they show that correcting for dependence using $z(\hat{\rho})$ may still be a good practice, even for values of n as small as 10, because the presence of dependence is likely to have much stronger adverse effects on inference than deviations from joint normality. See also Costantini and Lupi (2013).

3.2 Data and descriptive statistics

The raw data in the meta-analysis consist of 288 (12 studies times 24 commodities) unit root test statistics, reported as indicated in the last column of Table 1. However, these are not comparable since they are based on different tests and different sample sizes. What are comparable are the associated p-values, which can be then used to compute also comparable z-scores.

Nonetheless, in applied econometrics it is a rare practice to report the p-values for unit root tests. This is so because, as it is widely known, under the unit root hypothesis the test statistics follow nonstandard distributions; hence, the test developers would typically report simulated percentiles of these null distributions for the practitioners to use as critical values in their empirical applications. The applied researcher would then report the test statistic computed in her particular sample along with some indication (a "star") that the null hypothesis has been rejected, after having compared the test statistic to the developer's critical values. It should not be surprising, therefore, that the p-values are not reported in the studies of our collection. The notable exception is Balagtas and Holt (2009), who do report bootstrapped p-values (24 cases). Fortunately, we have enough information from the reported results to map each test statistic to a p-value.

Firstly, Kellard and Wohar (2006) report the test statistics along with the 1%, 5% and 10% critical values obtained by bootstrapping for the LP test, and so does Ghoshray (2011) for the LS test. This amounts to 27 cases. In all of them, the computed test lies within the reported critical values, and thus the p-value can be approximated by log-interpolation (the critical values are linearly related to the the logarithm of the significance level). On the other hand, Leybourne et al. (2007) provide enough detailed information on the percentiles of the LKT distribution such that the 24 p-values in Ghoshray (2013) can also be interpolated.

Secondly, the remaining 264 p-values are obtained through Monte Carlo simulations. For each study i, each

test used in study i, and at each iteration, a random walk $\Delta y_t \sim N(0,1)$ with initial condition $y_0 \sim N(0,1)$ is generated for $t = 1, 2, \ldots, T_i$, where T_i is the sample size used in study i. Then, the relevant test statistic, τ_i, is computed as described in section 2.1, and stored. For tests performing a grid search over location parameters, we use $\lambda \in (0.1, 0.9)$. This process is repeated $M = 500,000$ times and the p-value is approximated with a crude frequency simulator, i.e., the proportion of times the reported test statistic is less than or equal to the simulated τ_i across the M repetitions.[6]

Figure 1 presents the distributions of the computed z-scores for each study, sorted chronologically. To ease visualization, the few z-scores less than -3.75 were replaced by -3.75, and the few z-scores greater than 1.1 were replaced by 1.1. It can be observed that all studies have a mixture of rejections and nonrejections and that, in general, the z-scores vary considerably. In addition, even though obscured by the variability of the results, there is a tendency for the z-scores of the most recent studies to concentrate in larger negative values of z. The median z-score of 9 studies lies on the rejection region, at a 10% significance level, 8 of them published after 2005 (beginning with KW).

On the other hand, the data are presented grouped by commodity in Figure 2. Commodities are sorted in ascending order by the value of $z(0)$. The heterogeneity across commodities is quite apparent, with the z values quite concentrated in cases such as timber, lead, lamb or copper, and widely spread such as in hides, wool and silver. A few outliers (11 out of 288 observations) are found, most of them associated with abnormally strong rejections of the unit root hypothesis. The median value of z lies in the 10% rejection region in all cases but in copper, banana, cocoa and silver.

3.3 Meta-regression

To describe the characteristics of the collected z-scores, and to learn about the factors driving their patterns, we consider the following meta-regression, for commodity $c = 1, 2, \ldots, 24$ and study $i = 1, 2, \ldots, 12$:

$$z_{ci} = \mu + \alpha_c + \gamma_1 \sqrt{T_i \cdot 2\bar{T}} + \gamma_2 T_i + \sum_{k=1}^{3} \beta_{J,k} J_{k,i} + \sum_{k=1}^{2} \beta_{R,k} R_{k,ci} + \cdots$$
$$\cdots + \sum_{k=1}^{3} \beta_{A,k} A_{k,ci} + \sum_{i=1}^{2} \beta_{B,k} B_{k,ci} + \sum_{k=1}^{3} \beta_{D,k} D_{k,ci} + \sum_{k=1}^{2} \beta_{P,k} P_{k,ci} + \varepsilon_{ci}, \quad (6)$$

where the z-score is regressed on a constant (the grand mean), a commodity-specific effect and a number of attributes of the studies and the unit root tests. The term ε_{ci} is a disturbance that measures modeling choices and other factors not captured by these attributes, and also captures any simulation or rounding errors that may have occurred in the computation of the p-values. The meta-regression features six groups of dummy variables (J_k, R_k, A_k, B_k, D_k and P_k) each of them capturing all possible mutually exclusive categories and thus saturating the regression (for instance, $J_1 + J_2 + J_3 = B_1 + B_2 = 1$). In order to avoid the "dummy variable trap", the associated coefficients are restricted to sum zero (for instance, $\beta_{J,1} + \beta_{J,2} + \beta_{J,3} = \beta_{B,1} + \beta_{B,1} = 0$). The same holds for the commodity effects, $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 0$. This representation has the advantage that the regression coefficients can be interpreted as effects relative to μ, the grand, unconditional mean.

Two study-specific characteristics are included. The first is the sample size, whose effects are captured by the term $\gamma_1 \sqrt{T_i \cdot 2\bar{T}} + \gamma_2 T_i$, where $\bar{T}$ is the average sample size across studies. This parametrization is made such that both terms, who entered the regression as a deviation from their sample averages, are of the same order of magnitude. Moreover, it eases the interpretation because, departing from $\bar{T}$, an increase in T_i will produce a change in z equal to $(\partial z_{ci}/\partial T_i)_{T_i=\bar{T}} = \gamma_1 + \gamma_2$.

The second study-specific characteristic is the main orientation of the journal where the studies are published,

[6] We verify that the procedures were correctly implemented by first replicating the critical values reported in the original unit root papers, and then by comparing the conclusions drawn by the computed p-values to the "stars" reported in the empirical studies. A Matlab program that replicates all these computations (the execution time is less than an hour with a typical desktop computer), along with the computed p-values are available as supplementary material to this paper.

measured by the dummy variables J_k. These variables account for the possibility of a publication bias (see, *inter alia*, Stanley and Jarrell, 1989). In particular, $J_1 = 1$ if the journal specializes in development economics (144 cases corresponding to the articles in the Journal of Development Economics or the Journal of International Development), $J_2 = 1$ if the journal is mainly about econometric methodology (96 cases published in either Studies in Nonlinear Dynamics and Econometrics, the Journal of Time Series Analysis or Econometrics Review), and $J_3 = 1$ for the remaining cases (48 in total, published in the American Journal of Agricultural Economics).

The remaining variables are specific to the unit root test, and thus vary with c and i. For the variables R_k, $R_1 = 1$ if the p-value was reported in the study (24 cases), whereas $R_2 = 1$ if it was computed either by simulation (213 cases) or interpolation (51 cases).

The next three groups capture attributes of the testing procedures. The dummy variables A_k categorize the alternative hypotheses entertained by the various unit root tests: $A_1 = 1$ if the alternative model is a TS process around a linear trend (135 cases corresponding to the ADF, NP and ERS tests); $A_2 = 1$, if it features structural breaks modeled by dummy variables (86 cases corresponding to the PER, ZA, LP, LS, PR and CKP tests); and $A_3 = 1$ for "flexible" alternatives (67 cases corresponding to the AB, EL KSS and LKT tests). On the other hand, the variables B_k control for differences in the null hypothesis: $B_1 = 1$ if no structural break is allowed under H_0 (228 cases), and $B_2 = 1$ if structural are allowed (60 cases corresponding to the PER, PR, LS, CKP and MLP tests). Finally, the variables D_k measure whether the test uses some form of detrending prior to testing, in order to increase power: $D_1 = 1$ if no detrending is used (158 cases corresponding to all "OLS" tests plus KSS); $D_2 = 1$ if it uses LM detrending (37 cases for all the "LM" tests); and $D_3 = 1$ if it uses GLS detrending (93 cases, all the "GLS" tests plus LKT).

Often, the z-scores are associated with statistics that are not always the result of a single test, but rather of a sequence of tests. For instance, León and Soto (1997) runs the ZA on those prices for whom the ADF test did not reject, whereas some of the tests in Kellard and Wohar (2006) and Ghoshray (2011) are only reported when they reject. Hence, the associated p-value is not equal to the probability of the null hypothesis, but instead to the probability of the null *given* a prior nonrejection or given that the test rejects. This may introduce a pretesting bias in favor of rejecting the null. Thus, in the meta-regression we introduce the dummy variables $P_1 = 1$ for direct tests and $P_2 = 1$ for pretesting procedures, to investigate whether the second group differs systematically from the first. We classify the following 112 cases as possibly biased: the PER test in Cuddington (1992), the ZA tests in León and Soto (1997), the LP tests in Kellard and Wohar (2006), the LS tests in Ghoshray (2011), the AB tests in Harvey et al. (2011), all tests in Ghoshray et al. (2014) and in Meng et al. (2017), and the EL tests in Winkelried (2018).

Table 2 shows OLS estimates of three variants of the meta-regression: (1) without controlling for the commodity-specific effects; (2) controlling for these effects; and (3) controlling for commodity effects and removing the 11 outliers in Figure 2. Robust standard errors clustered by commodities are reported in square brackets. Also, some evidence of cross-sectional dependence was found using the test advanced in Frees (1995), and so the Driscoll and Kraay (1998) standard errors, which are robust to the presence of cross-sectional dependence, are also reported, in curly braces. The estimated grand mean is -1.582 and -1.568 after controlling for commodity effects; the latter corresponds to a p-value of $\Phi(-1.568) = 0.058$. After removing outliers, this estimate increases slightly to -1.490, with an associated p-value of $\Phi(-1.490) = 0.068$. Thus, unconditionally, the tests tend to reject the nonstationarity hypothesis in commodity prices.

The point estimates of the remaining coefficients are similar across regressions, and the inferences are also robust to the way standard errors are computed. We focus next on the second set of estimates with Driscoll and Kraay (1998) standard errors. A first finding is that the study-specific characteristics play no role in explaining the z-scores. Even though the coefficients γ_1 and γ_2 are precisely estimated, capturing the curvature in the map from the unit root statistics to the z-scores, the sum $\gamma_1 + \gamma_2$, which approximates the effect of increasing the sample size, is not statistically different from zero. Thus, *ceteris paribus*, the z-scores from studies that use the most recent updates of the GY data set are not statistically different from those reported in the earliest contributions. In addition, we find no evidence of publication bias, as the effects of the main orientation of

the journal where the studies are published (β_J) appear to be insignificant. This is also true for the effects of the way the test results are reported (β_R).

In contrast, the characteristics of the unit root tests display significant effects, all of them with expected signs: tests that allow for more flexible alternative hypotheses tend to reject more often (lower z-scores), whereas those that add flexibility to the null hypothesis tend to reject less often (higher z-scores). Regarding the nature of the alternative hypothesis, the tests that use the basic linear TS specification as the alternative model are associated with an average z-score of $\mu + \beta_{A,1} = -1.568 + 0.503 = -1.065$ and a combined p-value of 0.143, whereas those that use more flexible alternatives are associated with an average z-score of $\mu + \beta_{A,3} = -1.568 - 0.211 = -1.779$ and a p-value of 0.038. The tests that incorporate broken linear trends are also associated with lower z-scores, $\mu + \beta_{A,2} = -1.568 - 0.292 = -1.860$, but the differences with respect to the grand mean are not statistically significant. Similarly, the average z-score of tests that allow and do not allow for structural breaks in the null hypothesis are, respectively, $\mu + \beta_{B,2} = -1.568 + 0.528 = -1.040$ (p-value of 0.149) and $\mu + \beta_{B,1} = -1.568 - 0.528 = -2.096$ (p-value of 0.018).

Regarding the usage of data detrending in order to improve power, it is found that more powerful tests are associated with more frequent rejections of the null hypothesis. This is a manifestation of the documented theoretical results that, in general, unit root tests lack statistical power. The average z-score for OLS tests (no pretreatment of the data) is $\mu + \beta_{D,1} = -0.782$ (p-value of 0.217), and amounts to $\mu + \beta_{D,2} = -2.104$ (p-value of 0.018) for LM tests and $\mu + \beta_{D,3} = -1.818$ (p-value of 0.035) for GLS tests. This finding, together with the lack of statistical significance of the effects of an increase in the sample size, suggests that the chronological tendency towards rejection reported in Figure 1 is due to the adoption of more powerful and flexible methods for unit root testing, rather than a mere increase in the available amount of information used for this purpose.

A final finding of relevance is that the effects of pretesting are not statistical significant (β_P). Thus, after controlling for a variety of other characteristics, the meta-regression indicates that the z-scores of direct tests behave similarly than those that are result of sequential procedures.

3.4 Combined evidence

Table 3 presents the main results of our analysis, i.e. the combined z-scores and the inferences drawn from them for each of the 24 commodity prices under consideration. Together with a set of baseline results (in the panel "Raw data") we also report calculations after removing the 11 outliers shown in Figure 2. In the table, the rows are sorted by the value of $z(0)$ so, under independence, the commodities at the top of the table are those who present the strongest evidence against a unit root (i.e., the strongest rejections of the null), and this evidence weakens as we move to the bottom of the table. As expected, under independence, the unit root is rejected in all cases. Nonetheless, the (relatively weak) evidence of cross-sectional dependence put forward by the Frees (1995) test in the meta-regressions suggests that this assumption may not be valid, calling for an extension of the analysis that allow for dependence among z-scores.

Regarding the upper bounds for ρ, consider the one at a 10% level of significance, $\bar{\rho}_{10}$. This bound is above one in 18 out of 24 cases, meaning that the null hypothesis will be rejected in all these cases regardless of the actual value of ρ. The bound is moderate (above 0.5) in 3 of the remaining cases, and small (below 0.4) in the last 2 cases. In the latter the unit root would not be rejected for small values of ρ, even though rejection occurs under independence. The bound at the 5% level of significance, $\bar{\rho}_5$, is more demanding: it is greater than one in 10 cases, moderate (above 0.5) in 8 cases, and small (below 0.40) in 5 cases. Even though a definite conclusion depends on the values of ρ the researcher considers appropriate for this application, it is quite apparent that the combined evidence points towards the rejection of the unit root hypothesis in the vast majority of cases.

On the other hand, the point estimate of ρ, $\hat{\rho}$, was very close to zero for 15 commodities, including cocoa and silver at the bottom of the list. For this reason, we adopted the conservative strategy, likely to be biased towards nonrejection of the null, of using the combined z-score $z(\hat{\rho}_*)$, where $\hat{\rho}_*$ is the the maximum between the commodity-specific correlation and the average correlation across all commodities (equal to 0.11). With

this, the unit root is strongly rejected in 21 cases, cannot be rejected at a 5% significance level (a weak nonrejection) in 2 cases (banana and silver) and cannot be rejected at a 10% significance level (a strong nonrejection) only in the case of copper. It should not be surprising to verify that nonrejections occurs in the cases where the upper bounds were close to the point estimates of ρ; strictly speaking, when the bounds were contained within the (unreported) confidence intervals for ρ.

The results barely change when outliers are removed. The removal of outliers affects the value of $z(\hat{\rho}_*)$ because it changes $z(0)$ for those commodities that contained an outlying z-score (aluminum, beef, cocoa, coffee, lamb, lead, rubber and timber); and because, given our conservative strategy, it may increase the value of $\hat{\rho}_*$ for those commodities with small $\hat{\rho}$, since the average correlation increases from 0.11 to the still small value of 0.25. As a result, the combined z-scores $z(\hat{\rho}_*)$ in the second panel of Table 3 are closer to zero in all cases. Yet, this produces only two qualitative changes with respect to the baseline results, as the unit root cannot be rejected now at a 10% significance level for the cases of cocoa and silver. Thus, the unit root is strongly rejected in 20 cases, weakly rejected in the case of banana, and is not rejected only for cocoa, copper and silver.

In Table 4 we present further robustness checks. The analysis is repeated 12 times, each with 11 studies after removing one study at a time. For the first 20 commodities, the results are the same as in the baseline case: the combined p-values are less than 0.05 (and in most cases, less than 0.01). Similarly, for the case of banana, the combined p-value never exceeds 0.10, whereas in the case of copper the combined p-value is always greater than 0.10.

For the case of cocoa, the rejection found in the baseline results is driven by the inclusion of a few studies that strongly reject the null hypothesis. Remarkably, the combined p-value rockets from 0.026 to 0.184 when the Meng et al. (2017) study is excluded. A similar behavior is found for the case of silver, when the baseline p-value (0.055) almost triples when removing Balagtas and Holt (2009) and more than quadruples when Meng et al. (2017) is excluded. Thus, we may conservatively regard these fragile cases as nonrejections.

4 Closing remarks

Our meta-analysis has confirmed a trend in the results of unit root testing of real primary commodity prices: the DS hypothesis is rejected more often when the alternative model is flexible enough to characterize the behavior of a stationary but persistent series, or when the testing procedures have enhanced power properties. Furthermore, the combined evidence against unit roots in real primary commodity prices, at least as measured in the GY data set, is quite strong. Out of 24 commodities, the DS hypothesis is categorically rejected in 20 cases, and cannot be strongly rejected only in one instance. In the remaining three cases, the conclusion depends on the significance level used and on whether outliers are included or excluded from the computations. Thus, the unit root cannot be strongly rejected in at most 4 cases.

The findings of our meta-analysis provide empirical support to the claims in studies such as Deaton and Laroque (1992, 1996, 2003), that real primary commodity prices ought to be taken as mean-reverting, and the limited evidence on this prediction is due to the low power of unit root tests. Moreover, along these lines, the combined evidence subscribes the conclusion reached by authors such as Ardeni and Wright (1992), Deaton (1999), Cashin et al. (2002), Kellard and Wohar (2006), Ghoshray (2013) and Winkelried (2018), that it might be more productive to concentrate future research efforts in studying dynamic features in real primary commodity prices other than their stationarity. Among other topics, the presence on long-lasting cycles in these prices, the extent and causes of their unduly short-term volatility, the possibility of nonlinear dynamics, and their predictability in the medium-run.

References

Ardeni, P. G. and Wright, B. (1992). The Prebisch-Singer hypothesis: A reappraisal independent of stationarity hypotheses. *Economic Journal*, 102(413):803–812.

Arezki, R., Hadri, K., Loungani, P., and Rao, Y. (2014). Testing the Prebisch-Singer hypothesis since 1650: Evidence from panel techniques that allow for multiple breaks. *Journal of International Money and Finance*, 42(C):208–223.

Ayat, L. and Burridge, P. (2000). Unit root tests in the presence of uncertainty about the non-stochastic trend. *Journal of Econometrics*, 95(1):71–96.

Balagtas, J. V. and Holt, M. T. (2009). The commodity terms of trade, unit roots, and nonlinear alternatives: A smooth transition approach. *American Journal of Agricultural Economics*, 91(1):87–105.

Bleaney, M. and Greenaway, D. (1993). Long-run trends in the relative price of primary commodities and in the terms of trade of developing countries. *Oxford Economic Papers*, 4(3):349–363.

Carrion-i Silvestre, J. L., Kim, D., and Perron, P. (2009). GLS-based unit root tests with multiple structural breaks under both the null and the alternative hypotheses. *Econometric Theory*, 25(6):1754–1792.

Cashin, P., Liang, H., and McDermott, C. (2000). How persistent are shocks to world commodity prices? *IMF Staff Papers*, 47(2):177–217.

Cashin, P., McDermott, C. J., and Scott, A. (2002). Booms and slumps in world commodity prices. *Journal of Development Economics*, 69(1):277–296.

Cheng, L. and Sheng, X. S. (2017). Combination of "combinations of p values". *Empirical Economics*, 53(1):329–350.

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, 20(2):249–272.

Choi, I. (2015). *Almost All About Unit Roots. Foundations, Developments, and Applications*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.

Costantini, M. and Lupi, C. (2013). A simple panel-CADF test for unit roots. *Oxford Bulletin of Economics and Statistics*, 75(2):276–296.

Cuddington, J. T. (1992). Long-run trends in 26 primary commodity prices: A disaggregated look at the Prebisch-Singer hypothesis. *Journal of Development Economics*, 39(2):207–227.

Cuddington, J. T., Ludema, R., and Jayasuriya, S. A. (2008). Prebisch-Singer redux. In Lederman, D. and Maloney, W. F., editors, *Natural Resources: Neither Curse nor Destiny*, chapter 5, pages 103–140. Standford Univesity Press.

Cuddington, J. T. and Urzua, C. M. (1989). Trends and cycles in the net barter terms of trade: A new approach. *Economic Journal*, 99(396):426–442.

Deaton, A. (1999). Conmmodity prices and growth in Africa. *Journal of Economic Perspectives*, 13(3):23–40.

Deaton, A. and Laroque, G. (1992). On the behaviour of commodity prices. *Review of Economic Studies*, 59(1):1–23.

Deaton, A. and Laroque, G. (1996). Competitive storage and commodity price dynamics. *Journal of Political Economy*, 104(5):896–923.

Deaton, A. and Laroque, G. (2003). A model of commodity prices after Sir Arthur Lewis. *Journal of Development Economics*, 71(2):289–310.

Demetrescu, M., Hassler, U., and Tarcolea, A.-I. (2006). Combining significance of correlated statistics with application to panel data. *Oxford Bulletin of Economics and Statistics*, 68(5):647–663.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Driscoll, J. C. and Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4):549–560.

Elliot, G., Rothenberg, T. J., and Stock, J. H. (1996). Efficient tests for an autorregresive unit root. *Econometrica*, 64(4):813–836.

Enders, W. and Holt, M. T. (2012). Sharp breaks or smooth shifts? An investigation of the evolution of primary commodity prices. *American Journal of Agricultural Economics*, 94(3):659–673.

Enders, W. and Lee, J. (2012). The flexible Fourier form and Dickey-Fuller type unit root tests. *Economics Letters*, 117(1):196–199.

Frees, E. (1995). Assessing cross-sectional correlations in panel data. *Journal of Econometrics*, 69(2):393–414.

Ghoshray, A. (2011). A reexamination of trends in primary commodity prices. *Journal of Development Economics*, 95(2):242–251.

Ghoshray, A. (2013). Dynamic persistence of primary commodity prices. *American Journal of Agricultural Economics*, 95(1):153–164.

Ghoshray, A., Kejriwal, M., and Wohar, M. E. (2014). Breaks, trends and unit roots in commodity prices: A robust examination. *Studies in Nonlinear Dynamics and Econometrics*, 18(1):23–40.

Gouel, C. and Legrand, N. (2017). Estimating the competitive storage model with trending commodity prices. *Journal of Applied Econometrics*, 32(4):744–763.

Grilli, E. and Yang, M. C. (1988). Primary commodity prices, manufactured goods prices, and the terms of trade of developing countries: What the long run shows. *World Bank Economic Review*, 2(1):1–47.

Hartung, J. (1999). A note on combining dependent tests of significance. *Biometrical Journal*, 41(7).

Harvey, D. I., Kellard, N. M., Madsen, J. B., and Wohar, M. E. (2010). The Prebisch-Singer hypothesis: Four centuries of evidence. *Review of Economics and Statistics*, 92(2):367–377.

Harvey, D. I., Leybourne, S. J., and Taylor, A. M. R. (2011). Testing for unit roots and the impact of quadratic trends, with an application to relative primary commodity prices. *Econometric Reviews*, 30(5):514–547.

Iregui, A. and Otero, J. (2013). The long-run behaviour of the terms of trade between primary commodities and manufactures: A panel data approach. *Portuguese Economic Journal*, 12(1):35–56.

Kapetanios, G., Shin, Y., and Snell, A. (2003). Testing for a unit root in the nonlinear STAR framework. *Journal of Econometrics*, 112(2):359–379.

Kellard, N. M. and Wohar, M. E. (2006). On the prevalence of trends in primary commodity prices. *Journal of Development Economics*, 79(1):146–167.

Kim, T.-H., Pfaffenzeller, S., Rayner, T., and Newbold, P. (2003). Testing for linear trend with application to relative primary commodity prices. *Journal of Time Series Analysis*, 24(5):539–551.

Lee, J. and Strazicich, M. C. (2003). Minimum Lagrange Multiplier unit root test with two structural breaks. *The Review of Economics and Statistics*, 85(4):1082–1089.

León, J. and Soto, R. (1997). Structural breaks and long-run trends in commodity prices. *Journal of International Development*, 9(3):347–366.

Leybourne, S., Kim, T.-H., and Taylor, A. R. (2007). Detecting multiple changes in persistence. *Studies in Nonlinear Dynamics and Econometrics*, 11(3):1–34.

Lumsdaine, R. L. and Papell, D. H. (1997). Multiple trend breaks and the unit-root hypothesis. *Review of Economics and Statistics*, 79(2):212–218.

Mariscal, R. and Powell, A. (2014). Commodity price booms and breaks: Detection, magnitude and implications for developing countries. Working Paper 444, InterAmerican Development Bank.

Meng, M., Lee, J., and Payne, J. E. (2017). RALS-LM unit root test with trend breaks and non-normal errors: Application to the Prebisch-Singer hypothesis. *Studies in Nonlinear Dynamics and Econometrics*, 21(1):31–45.

Newbold, P., Pfaffenzeller, S., and Rayner, A. (2005). How well are long-run commodity price series characterized by trend components? *Journal of International Development*, 17(4):479–494.

Ng, S. and Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69(6):1519–1554.

Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57(6):1361–1401.

Perron, P. and Rodriguez, G. (2003). GLS detrending, efficient unit root tests and structural change. *Journal of Econometrics*, 115(1):1–27.

Pfaffenzeller, S., Newbold, P., and Rayner, A. (2007). A short note on updating the Grilli and Yang commodity price index. *World Bank Economic Review*, 21(1):151–163.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Prebisch, R. (1950). *The Economic Development of Latin America and its Principal Problems*. United Nations Publications, New York.

Said, E. S. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Schmidt, P. and Phillips, P. C. B. (1992). LM tests for a unit root in the presence of deterministic trends. *Oxford Bulletin of Economics and Statistics*, 54(3):257–287.

Singer, H. W. (1950). The distribution of gains between investing and borrowing countries. *American Economic Review*, 40(2):473–485.

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives*, 15(3):131–150.

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

Winkelried, D. (2016). Piecewise linear trends and cycles in primary commodity prices. *Journal of International Money and Finance*, 64:196–213.

Winkelried, D. (2018). Unit roots, flexible trends, and the Prebisch-Singer hypothesis. *Journal of Development Economics*, 132:1–17.

Zanias, G. P. (2005). Testing for trends in the terms of trade between primary commodities and manufactured goods. *Journal of Development Economics*, 78(1):49–59.

Zivot, E. and Andrews, D. W. K. (1992). Further evidence on the great crash, the oil price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, 10(3):251–270.

Table 1. *Studies in the meta-analysis*

		Last	TS	Test	H_1	SC in H_0	Detrend	p	Source
Cuddington (1992)	Cud	1983	12	ADF	Linear trend	No	No	*b*	Table 1 (p. 213)
				PER	Broken trend (1)	Yes	No	*b*	Table 1 (p. 213)
León and Soto (1997)	LeSo	1992	20	ADF	Linear trend	No	No	*b*	Appendix 1 (p. 366)
				ZA	Broken trend (1)	No	No	*b*	Table 2 (p. 354)
Kim et al. (2003)	KPRN	1998	8	ADF	Linear trend	No	No	*b*	Table III (p. 545)
Newbold et al. (2005)	NPR	2002	7	ADF	Linear trend	No	No	*b*	Table 2 (p. 483)
Kellard and Wohar (2006)	KW	1998	14	NP	Linear trend	No	GLS	*b*	Table 1 (p. 156)
				LP	Broken trend (2)	No	No	*c*	Table 2 (p. 157)
Balagtas and Holt (2009)	BH	2003	19	KSS	Nonlinear	No	No	*a*	Table 1 (p. 95)
Ghoshray (2011)	Gho11	2003	13	PR	Broken trend (1)	Yes	GLS	*b*	Table 2 (p. 247)
				LS	Broken trend (2)	Yes	LM	*c*	Table 3 (p. 248)
Harvey et al. (2011)	HLT	2003	16	ERS	Linear trend	No	GLS	*b*	Table 2 (p. 537)
				AB	Quadratic trend	No	GLS	*b*	Table 2 (p. 537)
Ghoshray (2013)	Gho13	2009	18	KLT	$I(0)$ regimes	No	GLS	*c*	Table 1 (p. 158)
Ghoshray et al. (2014)	GKW	2008	16	CKP	Broken trend (2)	Yes	GLS	*b*	Table 6 (p. 35)
Meng et al. (2017)	MLP	2007	21	MLP	Broken trend (2)	Yes	LM	*b*	Table 3 (p. 38)
Winkelried (2018)	Win	2010	20	ADF	Linear trend	No	No	*b*	Tables 1-4 (pp. 16-19)
				EL	Fourier trend	No	No	*b*	Tables 1-4 (pp. 16-19)

Notes: The figures in column "Last" are the last year in the sample; the effective sample size used in the unit root tests is, at most, T = Last − 1901. Column "TS" shows the number of real commodity prices, out of 24, found to be TS. Column "Test" lists the unit root tests used in each study, as discussed in section 2.1. The following three columns are characteristics of these tests: "H_1" shows the alternative hypotheses (in parentheses, the number of structural breaks considered, when applicable); "SC in H_0" indicates whether structural changes are allowed under H_0; and "Detrend" indicates if LM or GLS detrending is used to increase power. Finally, the column "p" refers to the way the p-values were computed, as described in section 3.2: (*a*) directly reported in the study; (*b*) by simulating the null distribution; or (*c*) by interpolation.

Table 2. *Meta-regressions*

	(1)		(2)			(3)		
μ Grand mean	−1.583	[0.119]***	−1.568	[0.131]***	{0.124}***	−1.534	[0.123]***	{0.112}***
γ_1	1.533	[0.478]***	1.563	[0.444]***	{0.708}**	1.509	[0.434]***	{0.712}**
γ_2	−1.592	[0.494]***	−1.620	[0.458]***	{0.738}**	−1.564	[0.448]***	{0.742}**
$\gamma_1 + \gamma_2$	−0.059	[0.039]	−0.057	[0.036]	{0.063}	−0.056	[0.034]	{0.063}
$\beta_{J,1}$ Development journals	0.113	[0.170]	0.047	[0.156]	{0.159}	0.041	[0.157]	{0.149}
$\beta_{J,2}$ Econometric journals	−0.046	[0.167]	−0.112	[0.155]	{0.221}	−0.056	[0.147]	{0.225}
$\beta_{J,3}$ Other journals	−0.067	[0.313]	0.065	[0.283]	{0.333}	0.015	[0.275]	{0.345}
$\beta_{R,1}$ Reported p-values	0.133	[0.172]	0.129	[0.181]	{0.188}	0.154	[0.179]	{0.184}
$\beta_{R,2}$ Computed p-values	−0.133	[0.172]	−0.129	[0.181]	{0.188}	−0.154	[0.179]	{0.184}
$\beta_{A,1}$ H_1: Linear trend	0.357	[0.194]*	0.503	[0.195]**	{0.238}**	0.499	[0.191]**	{0.245}*
$\beta_{A,2}$ H_1: Broken trend	−0.278	[0.265]	−0.292	[0.264]	{0.307}	−0.300	[0.257]	{0.318}
$\beta_{A,3}$ H_1: Flexible	−0.079	[0.238]	−0.211	[0.225]	{0.093}**	−0.199	[0.223]	{0.097}*
$\beta_{B,1}$ No breaks in H_0	−0.521	[0.248]**	−0.528	[0.229]**	{0.247}**	−0.515	[0.220]**	{0.266}*
$\beta_{B,2}$ Breaks in H_0	0.521	[0.248]**	0.528	[0.229]**	{0.247}**	0.515	[0.220]**	{0.266}*
$\beta_{D,1}$ OLS	0.825	[0.386]**	0.786	[0.400]*	{0.340}**	0.662	[0.418]	{0.336}*
$\beta_{D,2}$ LM	−0.563	[0.110]***	−0.536	[0.112]***	{0.080}***	−0.449	[0.124]***	{0.076}***
$\beta_{D,3}$ GLS	−0.262	[0.050]***	−0.250	[0.052]***	{0.026}***	−0.213	[0.059]***	{0.023}***
$\beta_{P,1}$ No pretesting	0.005	[0.147]	0.015	[0.138]	{0.193}	0.007	[0.125]	{0.207}
$\beta_{P,2}$ Pretesting	−0.005	[0.147]	−0.015	[0.138]	{0.193}	−0.007	[0.125]	{0.207}
Observations	288		288			277		
Adjusted R^2	0.191		0.258			0.208		
Commodity effects	No		Yes			Yes		
Outliers	Yes		Yes			No		
Frees (1995) test	0.357**		0.238*			0.279*		

Notes: Least squares estimations for three variants of equation (6): (1) baseline specification; (2) including commodity effects; (3) controlling for commodity effects and excluding outliers. Effects across categories are relative to the grand mean and sum to zero. Standard errors clustered by commodity in square brackets "[...]"; Driscoll and Kraay (1998) standard errors in curly brackets "{...}". * (**) [***] denotes rejection (H_0 : zero coefficient, or zero correlations in the Frees (1995) test of cross-sectional dependence) at a 10% (5%) [1%] significance level.

Table 3. *Combined tests*

		Raw data					Outliers correction				
	TS	$z(0)$	$\bar{\rho}_5$	$\bar{\rho}_{10}$	$\hat{\rho}_*$	$z(\hat{\rho}_*)$	$z(0)$	$\bar{\rho}_5$	$\bar{\rho}_{10}$	$\hat{\rho}_*$	$z(\hat{\rho}_*)$
Zinc	12	−9.99	≥ 1	≥ 1	0.48	−4.07 (0.000)	−9.99	≥ 1	≥ 1	0.48	−4.07 (0.000)
Rice	9	−8.67	≥ 1	≥ 1	0.11	−5.89 (0.000)	−8.67	≥ 1	≥ 1	0.25	−4.45 (0.000)
Timber	12	−7.92	≥ 1	≥ 1	0.56	−2.95 (0.002)	−6.37	≥ 1	≥ 1	0.74	−2.30 (0.011)
Sugar	11	−7.89	≥ 1	≥ 1	0.42	−3.34 (0.000)	−7.89	≥ 1	≥ 1	0.42	−3.34 (0.000)
Maize	9	−7.86	≥ 1	≥ 1	0.11	−5.34 (0.000)	−7.86	≥ 1	≥ 1	0.25	−4.03 (0.000)
Palmoil	9	−7.12	≥ 1	≥ 1	0.11	−4.84 (0.000)	−7.12	≥ 1	≥ 1	0.25	−3.65 (0.000)
Wheat	8	−6.12	≥ 1	≥ 1	0.11	−4.16 (0.000)	−6.12	≥ 1	≥ 1	0.25	−3.14 (0.001)
Rubber	10	−5.97	≥ 1	≥ 1	0.37	−2.65 (0.004)	−5.08	0.85	≥ 1	0.55	−1.99 (0.023)
Aluminum	9	−5.80	≥ 1	≥ 1	0.28	−2.85 (0.002)	−4.87	0.78	≥ 1	0.47	−2.04 (0.021)
Hides	9	−5.79	≥ 1	≥ 1	0.11	−3.94 (0.000)	−5.79	≥ 1	≥ 1	0.25	−2.97 (0.001)
Lead	9	−5.53	0.94	≥ 1	0.48	−2.20 (0.014)	−4.75	0.74	≥ 1	0.61	−1.79 (0.037)
Wool	8	−5.23	0.83	≥ 1	0.11	−3.55 (0.000)	−5.23	0.83	≥ 1	0.25	−2.68 (0.004)
Lamb	10	−5.12	0.79	≥ 1	0.32	−2.41 (0.008)	−4.90	0.87	≥ 1	0.62	−1.91 (0.028)
Beef	6	−4.80	0.68	≥ 1	0.11	−3.26 (0.001)	−3.85	0.45	0.80	0.25	−2.05 (0.020)
Tea	7	−4.54	0.60	≥ 1	0.11	−3.08 (0.001)	−4.54	0.60	≥ 1	0.25	−2.33 (0.010)
Coffee	7	−4.49	0.59	≥ 1	0.11	−3.05 (0.001)	−4.10	0.58	≥ 1	0.37	−1.97 (0.024)
Cotton	7	−4.47	0.58	≥ 1	0.11	−3.04 (0.001)	−4.47	0.58	≥ 1	0.25	−2.29 (0.011)
Tin	7	−4.47	0.58	≥ 1	0.11	−3.04 (0.001)	−4.47	0.58	≥ 1	0.25	−2.29 (0.011)
Jute	6	−4.02	0.45	0.80	0.11	−2.73 (0.003)	−4.02	0.45	0.80	0.25	−2.06 (0.020)
Tobacco	8	−3.46	0.31	0.57	0.11	−2.35 (0.009)	−3.46	0.31	0.57	0.25	−1.77 (0.038)
Copper	5	−3.36	0.29	0.53	0.62	−1.21 (0.114)††	−3.36	0.29	0.53	0.62	−1.21 (0.114)††
Banana	4	−2.99	0.21	0.40	0.23	−1.59 (0.056)†	−2.99	0.21	0.40	0.25	−1.53 (0.063)†
Cocoa	3	−2.85	0.18	0.36	0.11	−1.94 (0.026)	−1.84	0.03	0.11	0.25	−0.98 (0.164)††
Silver	4	−2.35	0.09	0.21	0.11	−1.60 (0.055)†	−2.35	0.09	0.21	0.25	−1.21 (0.114)††

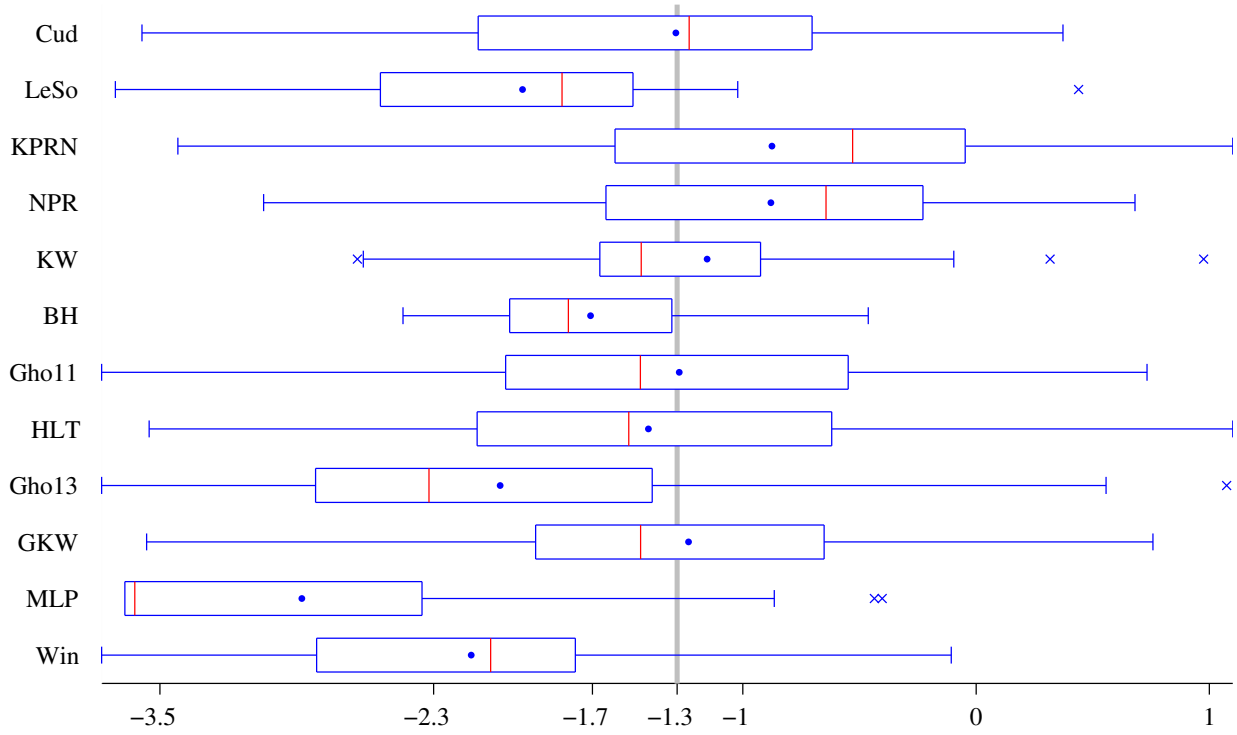
Notes: TS is the number of studies that reject the unit root hypothesis at a 10% significance level. The rows are sorted by the values of $z(0)$, the combined z-score that assumes independence. $\bar{\rho}_h$ $\{h = 5, 10\}$ is the minimum value of ρ such that the unit root hypothesis is rejected at an h% significance level, i.e., $z(\bar{\rho}_5) = -1.65$, $z(\bar{\rho}_{10}) = -1.282$ and $z(\rho) < z(\bar{\rho}_h)$ for all $\rho < \bar{\rho}_h$. $z(\hat{\rho}_*)$ allows for dependent outcomes using the estimator $\hat{\rho}_*$, the maximum between the commodity-specific correlation and the overall correlation, as the estimator of ρ. The p-values $\Phi(z(\hat{\rho}_*))$ are reported in parentheses.
† [††] indicates nonrejection at a 5% [10%] significance level ($0.05 < p \leq 0.10$) [($p > 0.10$)].

Table 4. *Robustness check*

	None	Cud	LeSo	KPRN	NPR	KW	BH	Gho11	HLT	Gho13	GKW	MLP	Win
Zinc	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Rice	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Timber	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.004	0.001	0.007	0.001
Sugar	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.003	0.001
Maize	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Palmoil	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Wheat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000
Rubber	0.004	0.003	0.002	0.002	0.002	0.002	0.004	0.002	0.003	0.005	0.005	0.039	0.002
Aluminum	0.002	0.002	0.001	0.003	0.001	0.001	0.001	0.003	0.001	0.001	0.002	0.037	0.001
Hides	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.005	0.001
Lead	0.014	0.017	0.010	0.010	0.012	0.011	0.015	0.049	0.014	0.010	0.010	0.010	0.013
Wool	0.000	0.000	0.002	0.000	0.000	0.001	0.001	0.002	0.002	0.003	0.000	0.009	0.003
Lamb	0.008	0.004	0.005	0.006	0.006	0.004	0.005	0.029	0.005	0.030	0.006	0.005	0.007
Beef	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.002	0.002	0.006	0.002	0.031	0.002
Tea	0.001	0.002	0.004	0.001	0.006	0.006	0.004	0.004	0.002	0.005	0.004	0.019	0.005
Coffee	0.001	0.004	0.004	0.003	0.003	0.003	0.004	0.004	0.003	0.005	0.004	0.033	0.006
Cotton	0.001	0.003	0.003	0.002	0.002	0.001	0.005	0.009	0.001	0.006	0.006	0.030	0.005
Tin	0.001	0.005	0.003	0.001	0.001	0.004	0.004	0.009	0.002	0.009	0.003	0.026	0.003
Jute	0.003	0.008	0.011	0.003	0.004	0.009	0.008	0.011	0.002	0.019	0.009	0.042	0.012
Tobacco	0.009	0.006	0.014	0.002	0.005	0.003	0.019	0.021	0.019	0.027	0.015	0.044	0.024
Copper	**0.114**††	**0.103**††	**0.124**††	**0.103**††	**0.107**††	**0.103**††	**0.152**††	**0.118**††	**0.122**††	**0.103**††	**0.104**††	**0.102**††	**0.138**††
Banana	**0.056**†	0.037	0.043	0.036	0.038	**0.054**†	0.042	**0.090**†	0.037	**0.094**†	**0.072**†	**0.081**†	**0.073**†
Cocoa	0.026	0.031	0.042	0.029	0.034	0.034	**0.068**†	0.043	0.033	0.029	0.040	**0.184**††	**0.073**†
Silver	**0.055**†	0.031	0.047	**0.070**†	**0.077**†	**0.127**††	**0.143**††	**0.063**†	**0.057**†	**0.108**††	**0.067**†	**0.242**††	**0.066**†

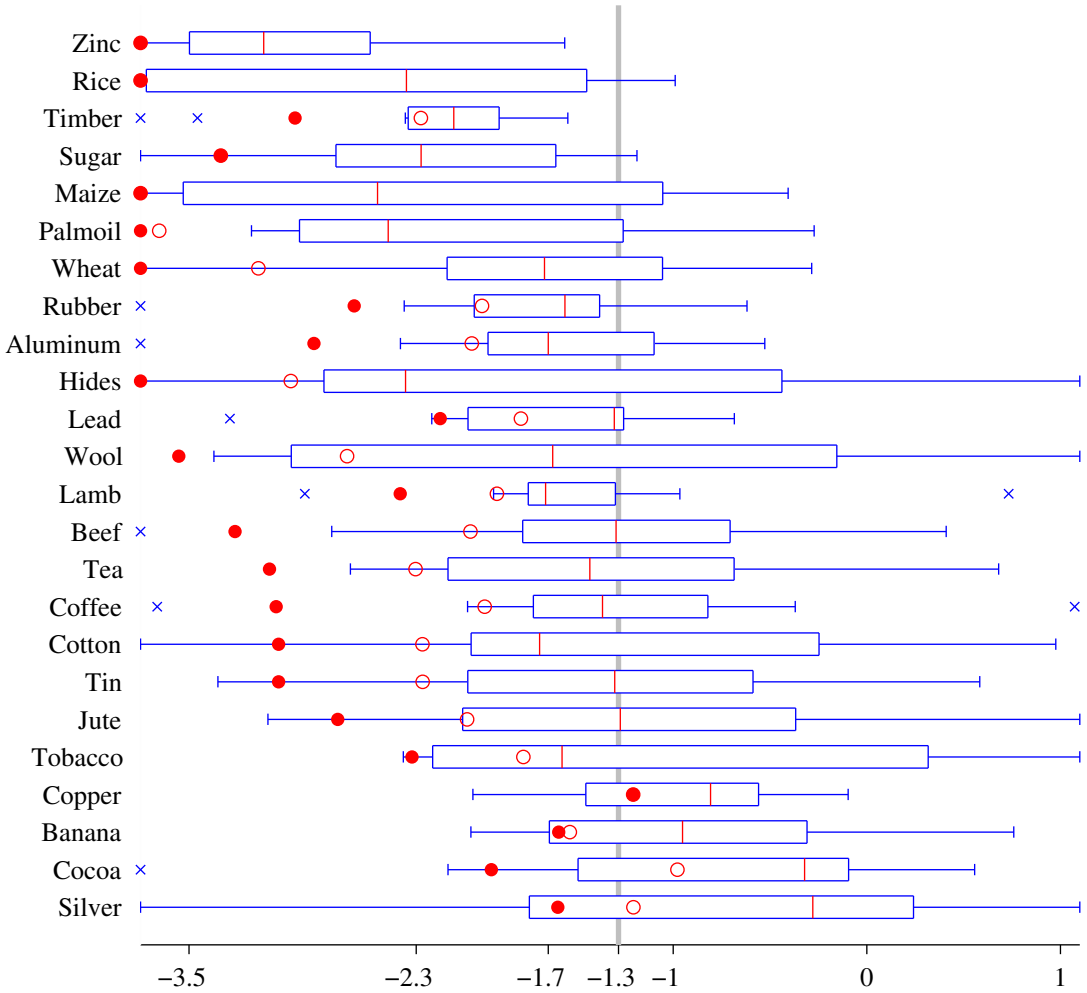
Notes: The figures are the combined p-values $\Phi(z(\hat{\rho}_*))$ computed after excluding one study (the columns) at a time, except for those of column "None" that shows the baseline results of Table 3. † [††] indicates nonrejection at a 5% [10%] significance level ($0.05 < p \leq 0.10$) [($p > 0.10$)].

Figure 1. *Computed z-scores across commodity prices, by study*



Notes: Standard Tukey's boxplots of the z-scores (the horizontal axis, the 10% critical value of -1.282 highlighted by a vertical gray line) in each study, for 24 commodity prices. Studies are sorted chronologically. The lines within the boxes are the median values of z; the filled circles, the mean values. "x" are outliers.

Figure 2. *Computed z-scores across studies, by commodity price*



Notes: Standard Tukey's boxplots of the z-scores (the horizontal axis, the 10% critical value of -1.282 highlighted by a vertical gray line) for each commodity price, across 12 studies. Commodity prices are sorted in increasing order according to the value of $z(0)$. The lines within the boxes are the medians of z. "x" are outliers. The filled circles mark the values of the combined z-score, $z(\hat{\rho})$; the empty circles, the combined z-score after removing outliers. These are reported in Table 3.