Econometrics **2014**, *2*, 72-91; doi:10.3390/econometrics2010072

OPEN ACCESS

econometrics

ISSN 2225-1146

www.mdpi.com/journal/econometrics

Article

Credible Granger-Causality Inference with Modest Sample Lengths: A Cross-Sample Validation Approach

Richard A. Ashley * and Kwok Ping Tsang

Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA; E-Mail: byront@vt.edu

* Author to whom correspondence should be addressed; E-Mail: ashleyr@vt.edu, Tel.: +1-540-231-6220, Fax.: 540-231-9288.

Received: 7 February 2014; in revised form: 14 March 2014 / Accepted: 18 March 2014 / Published: 25 March 2014

Abstract: Credible Granger-causality analysis appears to require post-sample inference, as it is well-known that in-sample fit can be a poor guide to actual forecasting effectiveness. However, post-sample model testing requires an often-consequential *a priori* partitioning of the data into an "in-sample" period – purportedly utilized only for model specification/estimation – and a "post-sample" period, purportedly utilized (only at the end of the analysis) for model validation/testing purposes. This partitioning is usually infeasible, however, with samples of modest length – e.g., $T \leq 150$ – as is common in both quarterly data sets and/or in monthly data sets where institutional arrangements vary over time, simply because there is in such cases insufficient data available to credibly accomplish both purposes separately. A cross-sample validation (CSV) testing procedure is proposed below which both eliminates the aforementioned *a priori* partitioning and which also substantially ameliorates this power versus credibility predicament – preserving most of the power of in-sample testing (by utilizing all of the sample data in the test), while also retaining most of the credibility of post-sample testing (by always basing model forecasts on data not utilized in estimating that particular model's coefficients). Simulations show that the price paid, in terms of power relative to the in-sample Granger-causality F test, is manageable. An illustrative application is given, to a re-analysis of the Engel and West [1] study of the causal relationship between macroeconomic fundamentals and the exchange rate; several of their conclusions are changed by our analysis.

Keywords: time series; Granger-causality; causality; post-sample testing; exchange rates

JEL classification: C18, C22, C52, F37

1. Introduction

The seminal contribution of [2] introduced the notion of "Granger-causality" and sparked a flurry of empirical implementations. In brief, the fluctuations in a time series x_t are said to Granger-cause fluctuations in a time series y_t if and only if an optimal forecasting model for y_t based on an otherwise-appropriately-wide information set, but omitting the past of x_t, forecasts y_t less well than an analogous model which additionally includes the past of x_t in the information set.[1]

Attention is usually restricted to linear forecasting models, in which restricted setting optimal modeling is relatively straightforward. This linearity assumption can itself be an issue, but it is the "appropriately-wide" information set restriction which can more easily be problematic. Indeed, this is the ultimate source of all examples in which the concept of Granger-causality yields apparently spurious results. It should be noted, however, that this problem with Granger-causality is essentially equivalent to the usual omitted-variables problem in econometric modeling, in which variables wrongly omitted from a model that are correlated with included ones lead to distorted inference on the included variables. Thus, Granger-causality testing merely calls on us to explicitly confront a problem which is endemic, but usually swept under the rug.

The initial spate of implementations (e.g., [4,5]) relied entirely on in-sample tests (usually just simple F-tests of the relevant model parameter restrictions) to infer whether or not the forecasting model for y_t over the wider information set is superior. Granger himself, however, soon became worried – as the result of observing a multitude of multivariate linear time series models which fit the sample data well, but forecast post-sample data very poorly – that these in-sample tests of causality were characteristically prone to distortion from "data mining." Such data mining is, of course, based on the fact that the model specification (variables and lag structure) is identified based on the same data used to fit and then to evaluate the model – e.g., see ([6]: p.281, 311). In essence, because we all tend to discard models which do not fit well, the fitted models we produce frequently fail to forecast well – or at all.[2] This concern led to the first post-sample implementations of Granger-causality – in [9,10] – testing explicitly whether or not the post-sample forecasts based on the wider information set are an actual improvement or not. A number of alternative tests for post-sample forecasting improvement – e.g. [11–19], among many

[1] See ([2]: p.430), although that paper mainly concentrates on a formulation based on spectral analysis; Breitung and Candelon[3] have continued this spectral strand of the analysis.

[2] As discussed in ([7]: Section 1) and ([8]: Chapter 7), this is the natural consequence of the fact that the fitting errors, whose size is being minimized by the estimation process itself, correspond to what Efron calls "apparent" rather than "true" errors.

others – were then developed in the ensuing years. [20] provides an up-to-date example describing and implementing a selection of the post-sample methods still popular.[3]

However – and despite one of us being an early and vocal advocate of post-sample testing – we must note that all post-sample implementations of the Granger-causality concept suffer from two inherent (and related) drawbacks:

(1) The analyst is always obliged to partition the available data set *at the outset* into an "in-sample" period – for use in identifying and estimating model specifications – and a "post-sample" or "holdout" period – which is to be reserved solely for evaluating which model provides superior forecasts. If done pristinely – *i.e.*, without looking at the data and at the forecasting performance of the models over various in-sample/post-sample splits – this partitioning is, at best, somewhat *ad hoc* and arbitrary.

(2) Post-sample Granger-causality testing tends to be feasible only where either the available data set is very long – so that a quite lengthy (and representative) post-sample period can be selected – or where the causal effect is so overwhelmingly strong as to hardly require statistical testing.

This first drawback has recently received renewed attention in [22] and in [23], both of which demonstrate the consequentiality of this choice for Granger-causation inference, and both of which therefore go on to propose statistical tests which are constructed so as to be robust to this choice. In their work one could say that the principal problem is actually an over-abundance of feasible in-sample/post-sample splits, leading to an awkwardly consequential nuisance parameter. The present work is distinct from theirs in that it is primarily aimed at settings in which the total amount of data available is modest – *i.e.*, where the principal problem is an overall paucity of data – which renders this sort of robustification problematic. Foreshadowing, our proposed procedure ameliorates this difficulty by both explicitly considering every possible in-sample/post-sample partitioning and by completely utilizing all of the scarce sample data, while always basing model forecasts on data not utilized in estimating that particular model's coefficients.

This data scarcity issue brings up the second drawback alluded to above. Using simulated data in an idealized setting, [24] showed that statistical testing for a mean square forecasting error improvement requires more data than one might expect. In particular, even with data generated from linear models with normally, identically, and independently distributed (NIID) error terms, one typically needs 80 to 100 post-sample observations in order to conclude that a 20% mean square error reduction is statistically significant at the 5% level. Basically, this is because one needs that much data in order to estimate a second moment (such as a mean squared error) with the requisite precision.[4]

---

[3] Over time it also became apparent – see [15,21] – that there is an important distinction to be made between choosing which model is closest to the the true (population) model versus which model provides the most accurate forecasts, especially for nested models. Despite the fact that the intuitive justification for the Granger-causation concept is grounded in forecasting, it is the former rather than the latter choice which is causation-relevant. The MSE-F post-sample test used in [20], takes this feature into account, however, so this particular aspect of post-sample testing for Granger-causation is not further discussed here. The MSE-F test is briefly described below in Section 3 below and also in [20]; see [14–16,19] for details.

[4] Even when, as here, the issue is relative predictability at the population level across two different information sets, [21] correctly argue that post-sample testing is inefficient; essentially, this is because it only uses a portion of the sample data

Additionally, another problem with a short post-sample period is that it can easily constitute a non-representative sample with regard to the putatively causing explanatory variables. For example, there might (or might not) be a strong and stable causal relationship between a variable y_t and lagged values of a variable x_t. But if there happens to be an unusually large (or small) amount of sample variation in x_t during the last portion of the data set, then a short post-sample model validation period can easily yield misleading results.

Thus, in settings where the available relevant data set is not very long – as is typically the case with quarterly macroeconomic data and as is more generally the case where institutional arrangements vary over time, so that only the most recent data are relevant – reliably informative statistical testing of the proposition that the post-sample mean square forecasting error from one model exceeds that of another is likely to require a post-sample period so lengthy as to leave insufficient in-sample data available for model identification and estimation.[5]

Section 2 introduces an elegant new Granger-causality test which – because it uses all of the available data at once in the testing procedure – both eliminates the need to decide *a priori* upon an in-sample versus post-sample partitioning of the available data set and also dramatically ameliorates the problems caused by a data set of modest length inducing the choice of a short post-sample period. Yet this new testing procedure retains a good deal of the credibility attached to post-sample testing, in that the relative performance of the estimated models used in the new test is always evaluated over data not used in the estimation of their coefficients; for this reason the new tests are denoted "cross-sample validation" or "CSV" Granger-causality tests below.

These CSV tests are new to the literature on Granger-causality with modest data sets, but the idea of estimating model coefficients over one part of the sample and then using them in another has long appeared in the statistics literature, where it is usually called "cross-validation". For example, [7] have independently proposed a model-comparison procedure which can be used to compare both cross-sectional and time series models; their approach "cross-validates" in a somewhat similar, but not identical, way to that of the CSV tests proposed here. (In particular, their procedure utilizes a large number of randomly-chosen sample-splits, whereas – as will be explained in Section 2 – the CSV tests explicitly examine every possible in-sample versus post-sample partitioning.)[6]

The results of calculations using simulated data to compare the empirical power of the new tests to that of the usual in-sample F test and to that of the MSE-F post-sample test are presented in Section 3 for sample lengths of 30, 60, and 120 periods. These results indicate that the power of the CSV Granger

available. On the other hand, this efficiency loss is empirically significant only where a lack of data forces one to specify a post-sample period which is short: it was probably not a very important factor in [20], for example, where a sample nearly 500 months in length allowed the authors to reserve 180 observations for post-sample testing.

5 The small sample lengths concentrated upon here could well be the natural consequence of having discarded a good deal of available sample data because one has statistically identified structural breaks in the data. In large-sample settings one might try to test for breaks and for Granger-causality all at once, as in [25]; see also [26].

6 The ([7]: p.8) procedure requires block bootstrap re-sampling when applied to serially dependent time series data, however, whereas the ordinary bootstrap will suffice for the tests proposed below. These authors cite several variations on the block bootstrap, including their preferred method, which is called geometric block bootstrapping.

causality tests proposed in Section 2 below is only modestly lower than that of the usual in-sample F test and (for post-sample periods of reasonable length) that their power is distinctly higher than that of the MSE-F post-sample test.

Note that this is the desired outcome – not a test with higher power than the in-sample F test. Rather, what the CSV Granger causality tests proposed here provide is a causality test with higher credibility than that of the in-sample F test, which credibility is obtained at a tolerable loss in power and while avoiding the problems (sample split arbitrariness, relatively low power, *etc.*) of the post-sample tests.

An illustrative application is given in Section 4, to a re-analysis of the [1] study of the causal relationship between macroeconomic fundamentals and the exchange rate. In their setting – with only 88 to 106 quarters of sample data available – post-sample Granger-causality testing was justifiably not considered feasible. Applying the cross-sample validation Granger-causality tests introduced here, we find that some of the Engel and West causality results are actually strengthened, but that the breadth of applicability of their conclusions is reduced. Section 5 concludes the paper.

2. A Cross-Sample Validation Test for Granger-Causality

For notational simplicity, we write the model for y_t over the full (unrestricted) information set in the usual multiple regression model format:

$$Y = X\beta^u + \varepsilon^u \tag{1}$$

where X is $T \times k$ and write the model for y_t over the restricted information set as:

$$Y = X^r\beta^r + \varepsilon^r \tag{2}$$

where the $T \times (k - g)$ array X^r is identical to X but omits the columns containing the data on the g putatively causative variables and where β^r omits the corresponding components. Here X might contain additional explanatory variables – even, for example, an error-correction term if y_t is co-integrated – as well as lagged values of y_t.

It is tacitly assumed here that the coefficient vector β^u is a constant over all T observations. this is assumed to have been assured either by having pruned the sample (which is why T might well be so modest in length) or by inclusion of appropriate explanatory variables at least approximately allowing for any structural changes within the data set.

Because the sampling distribution of the test statistic derived below is obtained using bootstrap simulation, Equation (1) must be specified with enough dynamics – e.g., lagged values of y_t and the other variables – that an assumption to the effect that the model errors (ε^u) are serially independent is tenable, but neither normality nor homoscedasticity needs to be assumed.

In Granger-causality analysis attention is usually restricted to linear models, but the right-hand sides of Equations (1) and (2) could instead have been nonlinear functions of the variables in the unrestricted and restricted information sets, respectively. Similarly, extending the present analysis

to multi-step forecasts (based on the unrestricted and restricted models) would be straightforward. The linear specifications used here greatly simplify the exposition which immediately follows in this section (and correspond to a very common assumption in this field) but, strictly speaking, inclusion of a sufficient number of lags of the dependent and explanatory variables in a linear regression model actually only suffices to ensure serial uncorrelatedness in the model errors, not the serial independence formally required for bootstrap simulation.[7] Absent extreme non-gaussianity in the errors, however, this distinction is, in and of itself, ordinarily of negligible significance, except insofar as linear model specifications reflect substantial model mis-specification in the form of wrongly omitted variables.

Indeed, it should be pointed out that an important implicit assumption in Equation (1) is that this specification includes the past values of all time series substantially relevant to the current value of y_t and especially any which are causally connected with the g time series omitted from X^r. This implicit assumption is both necessary and sufficient as to eliminate all of the usual counter-examples in which the Granger-causality concept itself becomes problematic, but it is nonetheless a strong assumption. On the other hand, it is also worth noting that this assumption is tacitly (and equally) made in any and all reduced-form regression modeling, so perhaps all that should be further mentioned here is that reasonable care must be taken (and common sense utilized) in specifying Equation (1).

Now suppose that the sample of T observations is split into two parts: the first τ observations and the remaining $T-\tau$ observations, where the value of τ is (for the moment) taken as given. Let the subscript "τ" denote an array consisting of just the first τ elements of the corresponding un-subscripted array and let the subscript "$-\tau$" similarly denote an array consisting of just the remaining $T-\tau$ elements.

Analogously, let $\hat{\beta}^u_\tau$ be the estimator of β^u in Equation (1) using only the first τ observations, let $\hat{\beta}^u_{-\tau}$ be the estimator of β^u in Equation (1) using only the last $T-\tau$ observations, and define $\hat{\beta}^r_\tau$ and $\hat{\beta}^r_{-\tau}$ similarly with regard to the estimators of β^r in the restricted regression, Equation (2). Clearly, $\hat{\beta}^u_\tau$ is simply $(X'_\tau X_\tau)^{-1} X'_\tau Y_\tau$ if (as would be likely for small T) OLS estimation is used, and similarly for $\hat{\beta}^u_{-\tau}$, $\hat{\beta}^r_\tau$, and $\hat{\beta}^r_{-\tau}$.[8]

Thus, if one takes the first τ observations to be the "in-sample" period and the remaining $T-\tau$ observations to be the "post-sample" period, then – without parameter updating – the $T-\tau$ post-sample forecasting errors made by the unrestricted model are just the $(T-\tau)\times 1$ array $\hat{\varepsilon}^u_{-\tau} \equiv Y_{-\tau} - X_{-\tau}\hat{\beta}^u_\tau$. Similarly, the post-sample forecasting errors made by the restricted model comprise the array $\hat{\varepsilon}^r_{-\tau} \equiv Y_{-\tau} - X^r_{-\tau}\hat{\beta}^r_\tau$.

Note, however, that one can also (in a completely analogous fashion) define the $\tau \times 1$ array of *"sample precasting"* errors $\hat{\varepsilon}^u_\tau \equiv Y_\tau - X_\tau \hat{\beta}^u_{-\tau}$, in which $\hat{\beta}^u_{-\tau}$ – the parameter estimator based on the (unrestricted model) data for the $T-\tau$ post-sample periods – is used to obtain the model errors for the first τ periods.

---

[7] Parametric nonlinear specifications for Equations (1) and (2) – and consequent CSV Granger causality analysis in that setting – are by no means ruled out, but would likely require substantially larger samples than are envisioned here. (It's not so much the fact that nonlinear least squares requires so much more data than does OLS, the problem is that the class of nonlinear models is so broad that the specification search process requires larger samples.) [27] provide a non-parametric in-sample Granger causality analysis framework, but effective non-parametric estimation requires even larger samples.

[8] One could imagine using EGLS, or GMM, or robust (LAD), or non-parametric estimation instead of OLS, but – in view of the small values of τ and $T-\tau$ envisioned here – their usefulness would be problematic.

Similarly, for the restricted model, the corresponding *"sample precasting"* errors are $\hat{\varepsilon}^r_\tau \equiv Y_\tau - X_\tau \hat{\beta}^r_{-\tau}$. In both cases these are just the prediction errors made in the first τ periods, using the parameter estimates obtained using only the data from the final $T - \tau$ periods.

For any given sample-split τ, one can easily compute both an unrestricted and a restricted sum of T squared "out-of-sample" prediction errors, $URSS_\tau$ and RSS_τ. Each of these is the sum of the squared prediction errors from all T periods in the data set, yet each is entirely based on errors made applying an estimated coefficient vector to explanatory variable data never used in its estimation. More explicitly:

$$URSS(\tau) = (\hat{\varepsilon}^u_\tau)'\hat{\varepsilon}^u_\tau + (\hat{\varepsilon}^u_{-\tau})'\hat{\varepsilon}^u_{-\tau} \tag{3}$$

$$= (Y_\tau - X_\tau \hat{\beta}^u_{-\tau})'(Y_\tau - X_\tau \hat{\beta}^u_{-\tau}) + (Y_{-\tau} - X_{-\tau}\hat{\beta}^u_\tau)'(Y_{-\tau} - X_{-\tau}\hat{\beta}^u_\tau) \tag{4}$$

and

$$RSS(\tau) = (\hat{\varepsilon}^r_\tau)'\hat{\varepsilon}^r_\tau + (\hat{\varepsilon}^r_{-\tau})'\hat{\varepsilon}^r_{-\tau} \tag{5}$$

$$= (Y_\tau - X^r_\tau \hat{\beta}^r_{-\tau})'(Y_\tau - X^r_\tau \hat{\beta}^r_{-\tau}) + (Y_{-\tau} - X^r_{-\tau}\hat{\beta}_\tau{}^r)'(Y_{-\tau} - X^r_{-\tau}\hat{\beta}_\tau{}^r) \tag{6}$$

In parsing the above equations, the reader is reminded that superscript "r" signifies that the g columns corresponding to the explanatory variables which putatively Granger-cause fluctuations in y_t have been removed, whereas subscript "τ" signifies that only the first τ rows corresponding to the first τ sample periods are used in the Y, X, and X^r arrays, and that the subscript "$-\tau$" signifies that only the last $T-\tau$ rows corresponding to the final $T - \tau$ sample periods are used in the Y, X, and X^r arrays.

Having obtained $URSS(\tau)$ and $RSS(\tau)$, the pseudo-F statistic

$$F_\tau \equiv \frac{\{RSS(\tau) - URSS(\tau)\}/g}{URSS(\tau)/(T-k)} \tag{7}$$

would be potentially useful in testing the null hypothesis that the coefficients on all g putatively Granger-causing explanatory variables are zero.

In practice, however, F_τ itself is of minimal interest, because it depends on the (arbitrary) sample-split at period τ. This dependence on the sample-split choice can be eliminated, though, by basing the Granger-causality inference on every possible value of τ. A straightforward way to do this is to utilize a sample quantile of the observed values of F_τ over all of the feasible values of τ as the test statistic.[9] More specifically, letting $\hat{Q}_\nu(x_1 \ldots x_m)$ denote the νth sample quantile of the distribution from which the

[9] The empirical power of tests based on the sample mean of the feasible F_τ values (and several variations involving un-equally weighted averages) was examined also, but the power of these tests is lower than that of tests based on the sample median. In addition, we naturally turned to quantile statistics because the distributions of the simulated F_τ are quite non-Gaussian.

observations $x_1 ... x_m$ are drawn – *i.e.*, the smallest value of x_i such that a fraction ν of $x_1 ... x_m$ do not exceed it – these sample order statistics can be expressed as:

$$\hat{Q}_\nu(F_{k+1} \ldots F_{T-k-1}) \tag{8}$$

where τ must lie in the interval $[k+1\,, T-k-1]$ so that both $\hat{\beta}^u_\tau$ and $\hat{\beta}^u_{-\tau}$ are computable. Thus, for example, $\hat{Q}_{0.50}$ is just the sample median of $F_{k+1} \ldots F_{T-k-1}$; $\hat{Q}_{0.75}$ is the sample third-quartile of $F_{k+1} \ldots F_{T-k-1}$; and $\hat{Q}_{1.00}$ is the maximum out of the values $F_{k+1} \ldots F_{T-k-1}$.

These sample order statistics, by construction, do not depend on τ. However – like F_τ itself – their finite-sample sampling distributions are unknown, even for conditionally homoscedastic model errors. Recalling that the raison d'être for the present approach is to obtain credible inference results despite the fact that the value of T is modest, the sample lengths envisioned here for use in calculating $\hat{Q}_\nu$ are inherently too small for the use of asymptotic results. Consequently, Granger-causation inferences based on $\hat{Q}_\nu$ must in practice be obtained using bootstrap methods, and results obtained in this way are quoted in Section 3 below.[10]

Granger-causality tests based on $\hat{Q}_\nu$ are aptly called "cross-sample validation' tests because they are based on applying the model coefficients estimated on one portion of the data to predicting the other portion of the data. Consequently, below we denote $\hat{Q}_{0.50}$ – the sample median of $F_{k+1} \ldots F_{T-k-1}$ – as the "$CSV50$" statistic. Analogously, we denote $\hat{Q}_{0.75}$ – the sample third-quartile of $F_{k+1} \ldots F_{T-k-1}$ – as the "$CSV75$" statistic, and so forth for the other values of ν.

Bootstrap inference ensures that the sizes of these cross-sample validation tests are reasonably accurate, even for the modest sample lengths considered here, but this needs to be checked. Further – compared to the power of the usual in-sample F test to detect Granger-causality – how large a price in power must one pay for the added credibility provided by these cross-validation tests? These issues are addressed in the next section.

3. Cross-Sample Validation Test Size and Power Comparisons Using Simulated Data

This section uses simulated data to compare the size and power of the cross-sample validation tests proposed above to that of both the usual in-sample F test and to that of a typical post-sample test. These results are designed to answer the following three questions:

---

[10] Bootstrap inference requires hardly more computer coding than does the sample evaluation of F_τ. And, using present equipment, bootstrap inference with $N_{boot} = 10,000$ simulations requires only 10 to 65 seconds of computer time as T varies from 30 to 120. Windows-based software is available from the authors which conveniently implements bootstrap-based Granger-causality inference based on $\hat{Q}_\nu$ for models such as Equation (1) (with $k <= 40$ and $T <= 4000$), optionally including multiple lags in the dependent variable and – where conditional heteroscedasticity in the model errors is a concern – using the "wild" bootstrap, as described in [28]. The reader should note, however, that, while not requiring $NIID$ model errors, the ordinary bootstrap still requires $\varepsilon^u \sim IID(0, \sigma^2)$ in Equation (1) and the wild bootstrap still requires serial independence in ε^u. As noted in Section 2, where linear modeling is sufficient then this serial independence can be ensured by including a sufficient number of lagged values of the dependent and explanatory variables in the specification of Equation (1).

- Are the bootstrapped cross-sample validation (CSV) tests well-sized in samples this small?
- Is the power of the cross-sample validation tests to detect Granger-causality close enough to that of the in-sample F test as to be a reasonable compromise?[11]
- For a reasonable post-sample forecasting period length, is the power of the cross-sample validation tests to detect Granger-causality so substantially higher than that of the post-sample test as to make this an attractive alternative?

More specifically, three kinds of test are considered:

1. The usual in-sample F test.[12] This test utilizes all T observations at once, with no sample split at all.
2. A standard post-sample test – in this case, the MSE-F test introduced in [14–16]. The MSE-F test statistic is:
$$MSE - F \equiv P \frac{\sum_{t=T-P+1}^{T} e_{r,t+1}^2 - e_{u,t+1}^2}{\sum_{t=T-P+1}^{T} e_{u,t+1}^2} \tag{9}$$
where P is the number of post-sample periods chosen, $e_{u,t+1}$ is the one-step-ahead forecasting error made by the unrestricted model in period t, and $e_{r,t+1}$ is the corresponding one-step-ahead forecasting error made by the restricted model. Both models are estimated using all data up to period t.[13]
3. And, finally, the cross-sample validation tests – based on the sample quantiles of $F_{k+1} \ldots F_{T-k-1}$ and embodied in test statistics such as $CSV50$, $CSV75$, and the like – as defined in Section 2 above.

After examining the size and power of the the tests in this section, we take up the issue as to whether the cross-sample validation tests can provide interestingly-distinct Granger-causality results in a practical setting in Section 4. Here the relative size and power of these three kinds of tests is compared using $M = 10,000$ artificially generated data sets, each of length of length T; results are given in Tables 1 and 2 for $T = 30$, 60, and 120.

---

[11] Recall that our CSV tests are not expected to have higher power than the in-sample F test: their added value (as with the post-sample tests) lies in their enhanced credibility.

[12] This is the standard test covered in most textbooks – e.g. ([29]: p.92).

[13] The notation used for the post-sample forecasting errors in Equation (9) is consistent with that of [19] – which defined analogous vectors of post-sample forecast errors, $\hat{\varepsilon}^u_{-(T-P)}$ and $\hat{\varepsilon}^r_{-(T-P)}$ – is intentionally distinct from that used in Equation (9) because the parameter estimates in the models used to obtain $e_{u,t+1}$ and $e_{r,t+1}$ are updated each period, whereas the out-of-sample prediction errors used in $\hat{\varepsilon}^u_{-(T-P)}$ and $\hat{\varepsilon}^r_{-(T-P)}$ are not.

Table 1. Rejection frequencies (empirical size) using data simulated from equation (10) (with coefficient on $x_{4,t}$ set to zero).

Test	T = 30	T = 60	T = 120
In-Sample F Test	0.0724	0.0602	0.0590
Cross-Sample Validation Tests:			
$\hat{Q}(0.00) - CSV00$	0.0521	0.0506	0.0510
$\hat{Q}(0.05) - CSV05$	0.0521	0.0487	0.0500
$\hat{Q}(0.10) - CSV10$	0.0516	0.0481	0.0430
$\hat{Q}(0.15) - CSV15$	0.0507	0.0452	0.0560
$\hat{Q}(0.20) - CSV20$	0.0472	0.0458	0.0560
$\hat{Q}(0.25) - CSV25$	0.0476	0.0456	0.0590
$\hat{Q}(0.30) - CSV30$	0.0473	0.0438	0.0570
$\hat{Q}(0.35) - CSV35$	0.0469	0.0463	0.0540
$\hat{Q}(0.40) - CSV40$	0.0442	0.0473	0.0510
$\hat{Q}(0.45) - CSV45$	0.0456	0.0467	0.0470
$\hat{Q}(0.50) - CSV50$	0.0464	0.0456	0.0500
$\hat{Q}(0.55) - CSV55$	0.0475	0.0451	0.0520
$\hat{Q}(0.60) - CSV60$	0.0477	0.0461	0.0540
$\hat{Q}(0.65) - CSV65$	0.0476	0.0486	0.0530
$\hat{Q}(0.70) - CSV70$	0.0482	0.0477	0.0510
$\hat{Q}(0.75) - CSV75$	0.0515	0.0457	0.0550
$\hat{Q}(0.80) - CSV80$	0.0520	0.0478	0.0540
$\hat{Q}(0.85) - CSV85$	0.0535	0.0459	0.0510
$\hat{Q}(0.90) - CSV90$	0.0549	0.0459	0.0570
$\hat{Q}(0.95) - CSV95$	0.0543	0.0490	0.0470
$\hat{Q}(1.00) - CSV100$	0.0543	0.0505	0.0570
Post-Sample MSE-F Tests:			
5 periods	0.0494	0.0463	0.0540
10 periods	0.0458	0.0449	0.0460
20 periods	0.0548	0.0435	0.0420
40 periods	-	0.0475	0.0470

These artificial data sets were generated from a dynamic multiple regression model of the form:

$$y_t = 0.7y_{t-1} + 0.2 + 0.3\,x_{1,t} + 0.3\,x_{2,t} + 0.0\,x_{3,t} + 0.3\,x_{4,t} + 0.0\,x_{5,t} + u_t \tag{10}$$

where u_t is generated as an NIID(0,1) variate for each observation in each data set.[14] As would be common, this regression model includes a lagged dependent variable and several explanatory control variables: $x_{1,t}$, $x_{2,t}$, and $x_{3,t}$, not all of which actually belong in the model. Equation (10) also includes two putatively causal variables: $x_{4,t}$ and $x_{5,t}$, one of which actually is causal. Aside from the lagged dependent variable, all of the explanatory variable values for each data set were generated (once) as AR(1) variates (with first-order autocorrelation of 0.50) and then held 'fixed in repeated samples' across all M artificial data sets.[15] The data on y_t for each artificial data set were then generated recursively from Equation (10).[16]

Equation (10) is typical, in size and kind, to the sorts of unrestricted models commonly used in Granger-causality analysis. In particular, this model is more general than the bivariate dynamic models used in the [1] study examined as an application in Section 4 below. Its assumption of NIID model errors seems reasonably innocuous since one might expect an analyst to include sufficient lagged terms in such a model as to eliminate any serial correlation in the errors and since the bootstrap inference used in implementing our method would in any case allow for any departures from normality and (using the wild bootstrap) from homoscedasticity.

The null hypothesis that neither $x_{4,t}$ nor $x_{5,t}$ Granger-causes y_t was then tested by applying all three kinds of test to a regression model (analogous to Equation (10)) which was fitted, using OLS, to each of these M data sets. Because exact sampling distributions are available for none of these tests with sample lengths this small, 5% critical points (and corresponding test rejection P-values for each artificial data set) were obtained using non-parametric bootstrap re-sampling to generate $N_{boot} = 10,000$ new T-samples based on this fitted regression model. More specifically, simulated values of y_1 ... y_T were obtained by recursion of this fitted model, using T "new" model errors generated by picking at random amongst the fitting errors.[17]

Two kinds of rejection frequency results (*i.e.*, estimates of the empirical power size and of the empirical power) for each of the three kinds of Granger-causality tests listed above – in each instance testing the null hypothesis that the coefficients on $x_{4,t}$ and $x_{5,t}$ are both zero – are collected in Tables 1 and 2 below for $M = 10,000$ artificial data sets of length $T = 30$, 60, or 120 generated in this way from Equation (10). The coefficient on the causal variable $x_{4,t}$ is set to zero for the size simulations in Table 1, so that the null hypothesis of no causality is correct for those simulations.

---

[14] Note that u_t is specified as NIID for these simulations, but the bootstrap simulations underlying the test proposed above require only that these model errors are serially independent.

[15] The empirical power results in Table 2 are not materially sensitive to re-generating these explanatory variable data using a different seed for the random number generator or for specifying differing levels of serial dependence in y_t or the explanatory variables.

[16] The value of y_0 for each data set was generated (also just once) as an independent unit normal variate.

[17] The explanatory variables which are not lags of the dependent variable – *i.e.*, $x_{1,t}$, ... $x_{5,t}$ in the present instance – are held fixed at their sample values across all of these bootstrap simulations also. As noted above, the Windows-based software implementing these bootstrap inferences (available from the authors) can handle more than one lag in the dependent variable and also allows one to choose whether the initial values of the lagged dependent variables in each bootstrap simulation are either set to the original sample values (the default) or picked at random from the sample data; the wild bootstrap is also optionally available for where conditionally heteroscedastic errors are a problem.

Table 2. Rejection frequencies (empirical power) using data simulated from equation (10).

Test	T = 30	T = 60	T = 120
In-Sample F Test	0.7726	0.9372	0.9998
Cross-Sample Validation Tests:			
$\hat{Q}(0.00) - CSV00$	0.0946	0.1133	0.1145
$\hat{Q}(0.05) - CSV05$	0.0946	0.1532	0.2588
$\hat{Q}(0.10) - CSV10$	0.1115	0.2583	0.4833
$\hat{Q}(0.15) - CSV15$	0.1319	0.3467	0.6481
$\hat{Q}(0.20) - CSV20$	0.1551	0.4166	0.7392
$\hat{Q}(0.25) - CSV25$	0.1920	0.4820	0.8083
$\hat{Q}(0.30) - CSV30$	0.2328	0.5303	0.8606
$\hat{Q}(0.35) - CSV35$	0.2648	0.5711	0.8944
$\hat{Q}(0.40) - CSV40$	0.2931	0.6063	0.9175
$\hat{Q}(0.45) - CSV45$	0.3203	0.6347	0.9391
$\hat{Q}(0.50) - CSV50$	0.3424	0.6571	0.9490
$\hat{Q}(0.55) - CSV55$	0.3672	0.6789	0.9587
$\hat{Q}(0.60) - CSV60$	0.3913	0.6925	0.9648
$\hat{Q}(0.65) - CSV65$	0.4147	0.7053	0.9695
$\hat{Q}(0.70) - CSV70$	0.4298	0.7204	0.9735
$\hat{Q}(0.75) - CSV75$	0.4327	0.7341	0.9759
$\hat{Q}(0.80) - CSV80$	0.4382	0.7429	0.9780
$\hat{Q}(0.85) - CSV85$	0.4232	0.7396	0.9803
$\hat{Q}(0.90) - CSV90$	0.4072	0.6948	0.9797
$\hat{Q}(0.95) - CSV95$	0.3820	0.5488	0.9494
$\hat{Q}(1.00) - CSV100$	0.3820	0.4471	0.5708
Post-Sample MSE-F Tests:			
5 periods	0.2574	-	-
10 periods	-	0.4959	-
20 periods	-	-	0.8296
40 periods	-	-	0.9280

The first thing to notice about Table 1 is that, for all three sample lengths, the empirical sizes of all of the bootstrapped tests – which is to say, for all of the tests other than the in-sample F test based on asymptotic theory – are all clustered right around 0.05. These results confirm that the bootstrap is both applicable and correctly implemented here. In contrast, the in-sample F test is noticeably oversized, at least for $T = 30$. This size distortion is is due to the fact that 30 observations is quite a small sample for

the use of asymptotic theory; the distortion would likely be larger if the entire sample were used for any sort of variable selection procedure.[18]

Table 2 in a few cases includes entries for post-sample tests with forecasting period lengths which are ludicrously small. For example, it is hardly credible that an analyst would truly sequester a post-sample period of length 10 or 20 periods from a total sample which is only 30 periods in length. On the other hand, it is interesting to at least look at the power of a test based on a five-period post-sample test in this case, so that entry is included in Table 2 nevertheless.

It is evident from the empirical power results in Table 2 that the in-sample F test has the highest power in each case. This result is to be expected: it is obviously helpful to estimate the model parameters using the entire data set; the object here is to obtain Granger-causality test results which are more convincing than those provided by the in-sample test because they do not rely for inference on the same sample data used to specify and estimate the models. Such credibility is to some extent provided by the post-sample MSE-F test, but the results in Table 2 indicate that this additional credibility comes at a high cost in terms of power. The cross-sample validation tests introduced here also provide higher-credibility Granger-causality inference than does the in-sample F test, but – in most cases – with substantially higher power than the post-sample tests.

Notably, the empirical power of the cross-sample validation tests based on the sample quantiles $\hat{Q}(\nu)$ with ν in the range [0.75, 0.90] is clearly the highest over the class of CSV tests. We interpret this to be the result of the implicit "trimming" in the sample quantiles of $F_{k+1} \ldots F_{T-k-1}$ for this range of ν values striking a balance between the higher informational content of the larger values of F_τ and their additional noisiness. In particular, we note that the empirical power of the $CSV100$, test – whose test statistic is $\sup F_\tau$ and is thus reminiscent of other $\sup F$ tests in the literature – is typically much smaller than the empirical power of the $CSV75$ test. Because a unique cross-sample validation test is desirable, our recommendation is to simply use the "third-quartile" or $CSV75$ cross-sample validation test in empirical applications.

The application given in the next section provides additional insights with regard to the relative merits of these different tests.

4. An Empirical Application: Do Fluctuations in Macroeconomic Fundamentals Granger-Cause Fluctuations in the Exchange Rate?

The standard view on the determination of a country's exchange rate is the asset-pricing model, in which the exchange rate is a function of the expected discounted values of future macroeconomic fundamentals – *i.e.*, cross-country differentials in output, money, interest rates, *etc*. But it is also a long-standing puzzle that models based on this theory have bleak empirical performance. In particular, exchange rates are well-approximated as random walks and do not appear to be forecastable using macroeconomic fundamentals. In an influential study, Engel and West [1], enhance the asset-pricing

[18] See also the discussion referenced in footnote 2. The $T = 120$ results for Table 1 are somewhat noisier and coarser-grained because only $M = 1,000$ simulations were used in this case.

model by adding two reasonable assumptions: that the subjective discount factor is close to one and that the macroeconomic fundamentals are highly persistent. Under these assumptions, the model predicts that exchange rates will behave like random walks and that innovations in exchange rates are correlated with news about future values of the macroeconomic fundamentals. They test their enhanced model by using quarterly data for six countries (with the U.S. as base, and a sample period of 1974 Q1 to 2001 Q3 in most cases) to look for Granger-causality between the growth rate in each country's exchange rate and the growth rate in each of several fundamental macroeconomic differential time series, relative to the U.S. The fundamentals variables Engel and West consider are, in their notation:[19]

- $mmd_t \equiv \triangle(m_t - m_t^*) \equiv$ money growth rate differential
- $ppd_t \equiv \triangle(p_t - p_t^*) \equiv$ inflation rate differential
- $ii_t \equiv (i_t - i_t^*) \equiv$ interest rate differential
- $iid_t \equiv \triangle(i_t - i_t^*) \equiv$ change in interest rate differential

According to Engel and West's exchange rate model, there should be no Granger-causality from these four time series to the exchange rate, but one should find Granger-causality from the exchange rate to these four fundamental variables. Because their sample is much too short to sequester a sub-sample period of reasonable length for post-sample testing, Engel and West use a likelihood ratio test (which is essentially equivalent to the usual in-sample F test) for their Granger-causality tests, including four lags of both the dependent and the independent variable in each model. Consistent with their model, they find almost no evidence for fluctuations in the fundamental variables Granger-causing fluctuations in the exchange rate; so these causal links are not considered here. But they are indeed able to find evidence for fluctuations in at least some of the fundamental variables Granger-causing fluctuations in the exchange rates for Germany, Italy, and Japan.[20] In particular, Engel and West are able to reject the null hypothesis of no-Granger-causality for the exchange rate at either the 5% or the 1% level of significance for ppd_t (in the case of Germany), for mmd_t and for ppd_t (in the case of Italy), and for all of mmd_t, ppd_t, ii_t, and iid_t (in the case of Japan).

But are these findings of Granger-causation merely artifacts due to the use of the same data in both estimating the bivariate relationships and in the causality testing? As noted above, the Engel and West data sets are too short for post-sample testing to be useful, but this is an excellent setting for the application of the cross-sample validation causality tests introduced here.

Table 3 summarizes the results. The sample lengths are given because several of the fundamental data series (for Italy and Japan) begin subsequent to 1974. Table 3 also indicates whether the results for this

---

[19] [1] also consider "yyd_t", comprising the growth rate in the difference between output in the specified country and that in the U.S. and "$mmyyd_t$", defined as mmd_t – yyd_t. These two fundamentals time series are not considered here because Engel and West did not find Granger-causality (significant at the 5% level on their in-sample test) between either of these variables and the exchange rate for any of the countries they considered.

[20] Engel and West's Table 3 also lists rejections at the 5% level of the null hypothesis of no Granger-causality from iid_t and from $mmyyd_t$ to the French exchange rate. These in-sample causality results were not examined here because the fitting errors for these two models – which regress iid_t or $mmyyd_t$ on their own past values and past values of the French exchange rate – both display very strong evidence of conditional heteroscedasticity and these two in-sample Granger-causality results disappeared when White-Eicker standard error estimates were used.

column were obtained using the usual bootstrap or using the wild bootstrap. The latter was necessary in four of the seven cases because the fitting errors of the underlying estimated regression model – for this country's exchange rate in terms of both its own past values and the past values of each fundamental variable – in these instances displayed severe conditional heteroscedasticity.[21] The next row of Table 3 displays the P-value at which the null hypothesis of no-Granger-causality can be rejected using the usual (in-sample) F test; as in Engel and West's work, all seven of these P-values remain less than 0.05 in these bootstrapped results.

Table 3. P-values for rejecting null hypothesis of no Granger-causality from macroeconomic fundamental to exchange rate using [1] data.

Country	**Germany**	**Italy**	**Italy**	**Japan**	**Japan**	**Japan**	**Japan**
Macroeconomic Fundamental	ppd_t	mmd_t	ppd_t	mmd_t	ppd_t	ii_t	iid_t
Sample Length	106	91	106	106	106	89	88
Bootstrap Type	ordinary	ordinary	wild	ordinary	wild	wild	wild
In-Sample F Test	0.014	0.060	0.032	0.024	0.007	0.013	0.001
Cross-Sample Validation 3^d Quantile Test:							
$\hat{Q}(0.75) - CSV75$	0.028	0.021	0.084	0.066	0.231	0.315	0.064
Post-Sample MSE-F Tests:							
5 periods	0.001	0.044	0.895	0.089	0.546	0.005	0.002
10 periods	0.030	0.048	0.850	0.097	0.047	0.016	0.011
20 periods	0.020	0.026	0.906	0.085	0.062	0.429	0.507
40 periods	0.772	0.143	0.947	0.067	0.052	0.620	0.683

The third-quartile cross-sample validation test – CSV75 – rejection P-values are given in the next row of Table 3.[22] These cross-sample validation test results are illuminating. First consider the evidence for Granger-causality running from ppd_t to the exchange rate. The evidence for this causal link in the data for Germany holds up quite well in the cross-sample validation tests. In contrast, the evidence for the analogous link in the data for Italy becomes substantially weaker; and the evidence for this ppd_t causal link disappears altogether in the data for Japan. Thus, the in-sample evidence in favor of ppd_t Granger-causing the exchange rate in the case of Germany (and, most likely, in the case of Italy) evidently represents the detection of an actual statistical regularity; in contrast, the in-sample evidence in the Japan case is apparently artifactual. The evidence for the causal link from mmd_t to the exchange rate in the data for Italy and Japan is still broadly present in the cross-sample validation CSV75 test results, although it is (as one might expect) a bit weaker for Japan than that provided by the in-sample test. Finally, the evidence for a causal link from ii_t or iid_t to the exchange rate in the data for Japan is again mixed: there

[21] In these four cases the conditional heteroscedasticity was so clearly visible in a time plot of the fitting errors that formal testing was beside the point. The use of robust standard error estimates in these models also yielded substantial changes in the in-sample F test P-values, with the ppd_t, ii_t, and iid_t rejections for Japan becoming notably more significant and the ppd_t rejection P-value for Italy rising from 0.004 to 0.033.

[22] Based on the empirical power results in Section 3, Table 3 focuses solely on the third-quartile cross-sample validation test results.

is still evidence (albeit somewhat weaker than from the in-sample result) for Granger-causality running from iid_t to the exchange rate, but the in-sample evidence for Granger-causality running from ii_t to the exchange rate appears to have been artifactual.

The remaining four rows of Table 3 display rejection P-values for the MSE-F tests with post-sample periods of lengths 5, 10, 20, and 40 quarters. It is unlikely that any analyst would employ these tests in samples this short; this point was mentioned above, but the reasoning underlying it deserves further comment here. Firstly, the smaller of these post-sample periods are hardly likely to constitute representative samples of the variation in the putatively causative explanatory variables; this un-representativeness will render the post-sample testing results erratic if the variation in these explanatory variables makes them either unusually influential or unusually non-influential within a brief post-sample period. (Also, even a modest amount of structural instability can be problematic with a brief post-sample testing period if one is "unlucky" in where a structural shift occurs.) Turning secondly to the longer (forty-quarter) post-sample period, this partitioning choice leaves very little data available for estimating the model coefficients over much of the post-sample testing.[23] That, of course, is why it is hardly credible that anyone would actually sequester this much data for post-sample testing with only 88 to 106 quarters of data available. That said, it is not surprising that the MSE-F post-sample rejection P-values vary erratically as the length of the post-sample period changes, typically becoming large for the forty-quarter post-sample period.

In summary, the cross-sample validation test results on the Engel and West data turn out to be quite illustrative: they are clearly distinct from the in-sample results, yet these more-credible results enrich – rather than broadly invalidate – Engel and West's original conclusions. In particular, the cross-validation results reinforce their contention that macroeconomic fundamentals can Granger-cause exchange rate fluctuations in some instances: certainly in the cases of ppd_t and the German exchange rate and of mmd_t and the Italian exchange rate – and probably also in the cases of ppd_t and the Italian exchange rate and the cases of both mmd_t and iid_t for the Japanese exchange rate. Yet several of Engel and West's causality inferences – ppd_t and ii_t for the Japanese exchange rate – turn out to be artifacts of the in-sample testing method used. Thus, our re-examination of Engel and West's data strengthens the empirical support for their exchange rate model in some countries, while also indicating that their in-sample causality analysis over-states the breadth of the evidence in support of the model's predictions.

5. Conclusions

"Credible" is one of those basic descriptors to which one does not give a precise statistical definition, but for which people nevertheless understand the meaning. For example, the reason that out-of-sample tests are more credible than in-sample tests is that, in practice, out-of-sample tests routinely yield fewer false rejections, even though both kinds of test have the same size. Maybe that is because it is so much easier – even for honest people – to invalidate the statistical size of an in-sample test, via data mining and

[23] Recall that the MSE-F tests are conducted with recursive parameter updating.

the like; maybe it is because post-sample forecasting cruelly exposes us to the effects of structural drift. The fact remains that estimated models almost always forecast less well in a post-sample period than they fit in the sample period. Therefore, most thoughtful analysts feel that post-sample testing results are more credible than in-sample ones. We assert here that our CSV tests are both much feasible to do in modest samples and – because our tests are not tied to any decision on the sample/post-sample split – more credible than post-sample testing, at a modest power penalty compared to the in-sample F test.

Post-sample Granger-causality testing still seems preferable for substantially large values of the sample length, T, as its efficiency loss will in such cases be out-weighed by its higher credibility.[24] But for small values of T – e.g., $T = 30$ to 60 – post-sample testing may be simply infeasible: here the data are so scarce as to make the assertion that the analyst "held back" even five observations from the model identification/estimation process risible. Also, even for somewhat larger values of T – e.g., $T = 60$ or $T = 120$ – a post-sample period of credible length might necessarily be so short as to provide little power in testing whether the post-sample forecasting errors of the restricted model are significantly larger than those of the unrestricted model. Or post-sample Granger-causality testing might easily yield erratic results in these settings, due to unusual sample variation in the putatively causing variables during the course of such a brief post-sample period.

In contrast, the usual in-sample F test utilizes all T observations to test the null hypothesis that the parameters on the putatively causative parameters are all zero, so one is in a notably better position to deal with a small data set. But this in-sample test is apt to routinely yield misleading Granger-causality inferences, simply because our modeling processes inherently pre-dispose us to find models which fit well. (After all, who among us has not found that our models tend typically to fit better than they forecast?) This tendency to systematically over-fit leads to in-sample detections of Granger-causation which is not actually present. Still, with samples of modest length, an analyst would heretofore have little choice but to utilize the in-sample test, despite this danger.

The "third-quartile" – $CSV75$ – cross-sample validation test proposed here resolves this predicament by utilizing all of the scarce sample data, while nevertheless always basing model predictions on coefficients estimated over data not used in making these predictions. Calculations based on simulated data indicate that this new test is well-sized and has empirical power not all that much lower than that of the less-credible in-sample test, so that the penalty paid for the additional confidence which can be accorded the results it provides is manageable. The empirical example based on the Engel and West (2005) data set, presented above in Section 4, illustrates the value of this new approach to Granger-causality analysis in settings where only modest amounts of sample data are available, especially in that several of their Granger-causality causality conclusions – e.g., for Japan – do not hold up under our cross-sample validation testing.

Finally, then, how does the present work indicate that modern Granger-causality analysis with a sample of modest length should be done?

---

[24] One might, however, for very large T turn instead to the method of [7].

- The first step should consist of a thoughtful specification of the unrestricted model for each variable – in $I(0)$ form – including (*i.e.*, conditioning upon) a reasonable approximation to the set of all importantly causative variables, an error-correction term (if cointegration is present), a sufficient number of lagged values of the dependent and other variables as to yield serially uncorrelated fitting errors, and whatever additional conditioning variables are necessary in order to remove obvious signs of structural drift (or shifts) during the data set. A plot of the model fitting errors is useful and warranted in this regard.
- Each of the restricted models is then specified and diagnostically checked in a similar fashion, as a nested model within the unrestricted model specification.
- The in-sample F test can then be applied to test for any particular causal link. If the null hypothesis (of no causality for this link) cannot be rejected at a culturally acceptable level of significance (5%, usually), then one can conclude that there is no real evidence for the existence of this causal link.
- If, in contrast, the null hypothesis of no causality is in fact rejected on the in-sample F test, then we suggest that a degree of skepticism is warranted: Could this result be an artifact of model mis-specification? Or – and what is usually essentially the same thing – are there non-homogeneities across the sample inducing a spurious inference? Or, is this rejection of the null hypothesis simply the result of an ordinary sampling fluctuation? To ameliorate, if not entirely relieve, these skepticism-inducing worries, we suggest the application of the CSV Granger causality test – typically $\hat{Q}(0.75)$ – as described above. If the null hypothesis of no causality is still rejected on the CSV tests, then it is reasonable to take this set of results as strong evidence in favor of the causal link actually existing. If, in contrast, the null hypothesis is no longer rejected on the CSV tests, then the initial skepticism – most especially with regard to model mis-specification and/or structural instability over the sample period – would seem to be warranted.

Acknowledgements

The authors wish to thank L. Kilian, M. McCracken, C. Parmeter, and K. D. West for helpful comments and we thank Scott Gilbert for both discussions and for his ancillary theoretical work on the CSV test. Further, we appreciate the feedback provided by three anonymous reviewers. The latest version of this paper is posted at http://ashleymac.econ.vt.edu/ashleyprofile.htm.

Author Contributions

Overall, both authors contributed equally to this project.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Engel, C.; West, K.D. Exchange Rates and Fundamentals. *J. Polit. Econ.* **2005**, *113*, 485–517.
2. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438.
3. Breitung, J.; Candelon, B. Testing for Short-and Long-run Causality: A Frequency-Domain Approach. *J. Econom.* **2006**, *132*, 363–378.
4. Sims, C.A. Money, Income, and Causality. *Am. Econ. Rev.* **1972**, *62*, 540–552.
5. Pierce, D.A.; Haugh, L.D. Causality in Temporal Systems: Characterizations and a Survey. *J. Econom.* **1977**, *5*, 265–293.
6. Granger, C.W.J; Newbold, P. *Forecasting Economic Time Series*; Academic Press: New York, USA, 1977.
7. Racine, J.S.; Parmeter, C. Data-Driven Model Evaluation: A Test for Revealed Performance. In *Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*; Ullah, A., Racine, J.S., Su, L., Eds.; Oxford University Press: Oxford, UK, 2013. Available online: http://www.ncsu.edu/cenrep/workshops/documents/modeval.pdf.
8. Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans*; Society for Industrial and Applied Mathematics: Philadelphia, USA, 1982.
9. Ashley, R.; Granger, C.W.J.; Schmalensee, R. Advertising and Aggregate Consumption: An Analysis of Causality. *Econometrica* **1980**, *48*,1149–1168.
10. Ashley, R. Inflation and the Distribution of Price Changes across Markets: A Causal Analysis. *Econ. Inq.* **1981**, *19*, 650–660.
11. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263.
12. West, K.D. Asymptotic inference about predictive ability. *Econometrica* **1996**, *65*, 1067–1084.
13. Ashley, R. A New Technique for Postsample Model Selection and Validation. *J. Econ. Dyn. Control* **1998**, *22*, 647–665.
14. Gilbert, S. Sampling Schemes and Tests of Regression Models. Manuscript. Department of Economics, Southern Illinois University at Carbondale, 2001.
15. Clark, T.; McCracken, M. Test of Equal Forecast Accuracy and Encompassing for Nested Models. *J. Econom.* **2001**, *105*, 85–110.
16. Clark, T.; McCracken, M. Evaluating Direct Multi-Step Forecasts. *Econom. Rev.* **2005**, *24*, 369–404.
17. Clark, T.; West, K. Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. *J. Econom.* **2006**, *135*, 155–186.
18. Clark, T.; West, K. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *J. Econom.* **2007**, *138*, 291–311.
19. McCracken, M.W. Asymptotics for out of sample tests of Granger causality. *J. Econom.* **2007**, *140*, 719–752.

20. Ashley, R.; Ye, H. On the Granger Causality between Median Inflation and Price Dispersion. *Appl. Econ.* **2012**, *44*, 4221–4238.
21. Inoue, A.; Kilian, L. In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econom. Rev.* **2004**, *23*, 371–402.
22. Hansen, P.R.; Timmermann, A. *Choice of Sample Split in Out-of-Sample Forecast Evaluation*; European University Institutute Working Papers ECO 2012/10; EUI: Fiesole, Italy, 2012; pp. 1–42.
23. Rossi, B.; Inoue, A. Out-of-Sample Forecast Tests Robust to the Choice of Window Size. *J. Bus. Econ. Stat.* **2012**, *30*, 432–453.
24. Ashley, R. Statistically Significant Forecasting Improvements: How Much Out-of-Sample Data is Likely Necessary? *Int. J. Forecast.* **2003**, *19*, 229–239.
25. Rossi, B. Optimal tests for nested model selection with underlying parameter instability. *Econom. Theory* **2005**, *21*, 962–990.
26. Pesaran, M.H.; Timmermann, A. Selection of Estimation Window in the Presence of Breaks. *J. Econom.* **2007**, *137*, 134–164.
27. Diks, C.; Panchenko, V. A New Statistic and Practical Guidelines for Nonparametric Granger Causality Testing. *J. Econ. Dyn. Control* **2006**, *30*, 1647–1669.
28. Gonçalves, S.; Kilian, L. Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *J. Econom.* **2004**, *123*, 89–120.
29. Davidson, R.; MacKinnon, J.G. *Estimation and Inference in Econometrics*; Oxford University Press: Oxford, UK, 1993.