Journal of Forecasting
J. Forecast. **24**, 453–464 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.960

A Note on In-Sample and Out-of-Sample Tests for Granger Causality

SHIU-SHENG CHEN*
Department of Economics, National Taiwan University, Taiwan

ABSTRACT

This paper studies in-sample and out-of-sample tests for Granger causality using Monte Carlo simulation. The results show that the out-of-sample tests may be more powerful than the in-sample tests when discrete structural breaks appear in time series data. Further, an empirical example investigating Taiwan's investment–saving relationship shows that Taiwan's domestic savings may be helpful in predicting domestic investments. It further illustrates that a possible Granger causal relationship is detected by out-of-sample tests while the in-sample test fails to reject the null of non-causality. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS Granger causality; out-of-sample forecast; forecast accuracy

INTRODUCTION

Granger causality tests have been used frequently to investigate the relationship between two or more variables of interest. Recent research, however, has re-examined Granger's original idea by arguing that the standard in-sample implementations of Granger's definition depart from the spirit of causality suggested by Granger (1969) in which 'causality' requires evidence of improved post-sample forecasting. In Ashley *et al*. (1980), they argue:

> . . . a sound and natural approach to such [Granger causality] tests must rely on the out-of-sample forecasting performance of models relating the original (non-prewhitened) series of interest.

This implies that out-of-sample forecasting performance may provide better information about causality.

Among studies attempting to reconsider Granger causality in the context of post-sample forecasts, McCracken (2004) derives the asymptotic distributions of several tests of equal forecast accuracy between two nested models. Chao *et al*. (2001) construct an out-of-sample test for Granger causality. Clark and McCracken (2001) provide some evidence on the asymptotic and finite-sample behaviour of tests of equal forecast accuracy and encompassing for nested models. However, a question remains to be answered: are in-sample or out-of-sample tests preferable? This is an important issue

* Correspondence to: Shiu-Sheng Chen, Department of Economics, National Taiwan University, No. 21 Hsu-Chow Road, 100 Taipei, Taiwan. E-mail: sschen@ntu.edu.tw

since people may find that there exist differences between in-sample and post-sample evidence in empirical studies (see Clark and McCracken, 2001; Awad and Goodwin, 1998). Further, Chao *et al*. (2001) report that an empirical illustration shows that the choice of in-sample versus out-of-sample Granger causality tests can crucially affect the conclusions about the predictive content. Recently, many empirical studies have begun to apply out-of-sample Granger causality tests. For instance, see Rapach and Weber (2002, 2004) and Lettau and Ludvigson (2001).

However, it has been shown in Clark and McCracken (2001) that the in-sample Granger causality test is more powerful than all of the out-of-sample tests they examine in the Monte Carlo simulations. This result is not surprising since the data generating processes they investigate are stationary. Does this result imply that the out-of-sample test is not useful in nested model selection? The answer may be negative. An important factor is neglected in the above simulation: the presence of underlying *parameter instability*.

Some recent papers start to consider the role of parameter instability under the context of out-of-sample forecasts. For instance, Rossi (2003) develops optimal tests of the joint null hypothesis of no (Granger) causality and parameter stability. Clark and McCracken (2003a) provide analytical and Monte Carlo evidence on how parameter instability affects tests of equal forecast accuracy and encompassing. They find that the instability lowers the power of out-of-sample forecast tests relative to in-sample tests. Those papers, however, consider only the structural breaks in the Granger-causal coefficients of interest. In contrast, Inoue and Kilian (2003) consider structural breaks in both Granger-causal coefficients and other non-Granger-causal coefficients, and conclude that the presence of structural changes in non-Granger-causal coefficients does not favour out-of-sample tests of predictability. Their conclusion, however, comes from a very small set of out-of-sample tests.[1]

We will examine whether the out-of-sample causality tests dominate in the presence of a specific setting of parameter instability. This paper is related to the existing literature in the following sense: on the one hand, as a complement to Rossi (2003) and Clark and McCracken (2003a), this paper focuses on structural breaks in non-Granger-causal coefficients. On the other hand, complementary to the work by Inoue and Kilian (2003), this paper investigates a larger set of out-of-sample tests. The experimental designs in this study are very close to the designs in Clark and McCracken (2003b), however, they focus on how to account for the weakness of the out-of-sample evidence on the Phillips curve relative to the in-sample evidence. Instead, this paper investigates the size and power properties of out-of-sample and in-sample tests for Granger causality.

The results from my Monte Carlo simulation confirm the findings in Inoue and Kilian (2003) that out-of-sample tests are not preferred when *F*-type equal MSE test and encompassing *t*-test are used. However, a new encompassing test proposed by Clark and McCracken (2001) outperforms in-sample tests when facing the presence of discrete structural breaks as discussed in this study. A two-stage procedure to deal with the causality test is then suggested and evaluated.

The rest of the paper is structured as follows. The following section reviews a brief comparison of in-sample and out-of-sample Granger causality tests. The third section presents the Monte Carlo simulation design. The results and discussion are reported in the fourth section. Then we provide a simple empirical example in the fifth section. A final section concludes.

[1] Inoue and Kilian (2003) consider only two out-of-sample tests: an *F*-type equal MSE test and an encompassing *t*-test proposed by Harvey *et al*. (1998).

IN-SAMPLE AND OUT-OF-SAMPLE GRANGER CAUSALITY TESTS

Consider the following nested models:

$$\text{Model 1} \quad y_t = \alpha_1 + \beta_{11} x_{1t-1} + u_{1t} \tag{1}$$

$$\text{Model 2} \quad y_t = \alpha_2 + \beta_{21} x_{1t-1} + \beta_{22} x_{2t-1} + u_{2t} \tag{2}$$

It is clear that model 2 nests model 1 under the non-Granger-causality null: $\beta_{22} = 0$. An out-of-sample Granger causality test means that the predictive ability of model 2 is better than that of model 1. For instance, if we use the mean square prediction error (MSPE) as a measure of prediction performance, then MSPE(2) < MSPE(1) implies that x_{2t-1} Granger causes y_t. In contrast, a conventional in-sample Granger causality test (GC test) simply uses an F-test for testing the null hypothesis: $\beta_{22} = 0$.

The out-of-sample forecast we consider here is a one-step-ahead prediction. The sample of observations $\{y_t, x_{1t}, x_{2t}\}_{t=1}^{T}$ is divided into in-sample and out-of-sample portions. There are R in-sample observations: $t = 1, \ldots, R$ and P out-of-sample observations: $t = R + 1, \ldots, R + P$. Obviously, $R + P = T$. A recursive scheme of estimation is used.

The following out-of-sample tests will be investigated and summarized in Table I:[2]

(a) The equal MSE tests: an F-type test (MSE-F) proposed by McCracken (2004), a t-test (MSE-T) proposed by Diebold and Mariano (1995), and the Granger and Newbold (1977) t-test (MSE-REG).
(b) The encompassing tests: a t-test (ENC-T) proposed by Harvey *et al*. (1998), a regression-based test (ENC-REG) proposed by Ericsson (1992), and a new test (ENC-NEW) developed by Clark and McCracken (2001).
(c) The Chao *et al*. (2001) test (CCS).

SIMULATION DESIGN

Our simulation design is simply a generalization of the design in Clark and McCracken (2001). Consider the following data generating process (DGP):[3]

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \alpha_t & \beta \\ 0 & \gamma_t \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_{y,t} \\ u_{x,t} \end{bmatrix}, \quad t = 1, \ldots, T \tag{3}$$

where $(u_{y,t}, u_{x,t})$ are i.i.d. standard normal random variables. It is worth noting that the coefficients (α_t, γ_t) are time-varying, which accounts for the parameter instability.

Letting $\theta_t = (\alpha_t, \gamma_t)'$, we consider the following two designs:

(DGP1) $\theta_t = \theta_0 \forall t$

[2] For the details of these tests, see Clark and McCracken (2001).
[3] Note that this study intends to expand on the experiments in Clark and McCracken (2001). The settings of parameter values follow Clark and McCracken (2001) with modifications of parameter instability. The DGPs are not economically motivated.

Table I. Out-of-sample tests

(a) The equal MSE tests

$$\text{MSE}-\text{T} = P^{1/2} \times \frac{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)}{\sqrt{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)^2}}$$

$$\text{MSE}-\text{F} = P \times \frac{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)}{P^{-1}\sum_t \hat{u}_{2,t+1}^2}$$

$$\text{MSE}-\text{REG} = P^{1/2} \times \frac{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)}{\sqrt{\left(P^{-1}\sum_t (\hat{u}_{1,t+1} - \hat{u}_{2,t+1})^2\right)\left(P^{-1}\sum_t (\hat{u}_{1,t+1} + \hat{u}_{2,t+1})^2\right) - \left(P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{2,t+1}^2\right)\right)^2}}$$

(b) The encompassing tests

Let $c_{t+1} = \hat{u}_{1,t+1}(\hat{u}_{1,t+1} - \hat{u}_{2,t+1})$ and $\bar{c} = P^{-1}\sum_t c_{t+1}$

$$\text{ENC}-\text{T} = P^{1/2} \times \frac{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{1,t+1}\hat{u}_{2,t+1}\right)}{\sqrt{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{1,t+1}\hat{u}_{2,t+1}\right)^2 - \bar{c}^2}}$$

$$\text{ENC}-\text{REG} = P^{1/2} \times \frac{P^{-1}\sum_{t=R}^{T} \left(\hat{u}_{1,t+1}^2 - \hat{u}_{1,t+1}\hat{u}_{2,t+1}\right)}{\sqrt{P^{-1}\sum_t (\hat{u}_{1,t+1} - \hat{u}_{2,t+1})^2 \left(P^{-1}\hat{u}_{1,t+1}^2\right) - \bar{c}^2}}$$

$$\text{ENC}-\text{NEW} = P \times \frac{P^{-1}\sum_t \left(\hat{u}_{1,t+1}^2 - \hat{u}_{1,t+1}\hat{u}_{2,t+1}\right)}{P^{-1}\sum_t \hat{u}_{2,t+1}^2}$$

(c) CCS test

$$\text{CCS} = P \times \overline{m}'\hat{\Omega}^{-1}\overline{m}$$

where $\overline{m} = P^{-1}\sum_t \hat{u}_{1,t+1}x_{2t}$

$$\hat{\Omega} = \left(P^{-1}\sum_t \hat{u}_{1,t+1}^2\right)\left(P^{-1}\sum_t x_{2t}x_{2t}'\right)$$

Note: $\hat{u}_{1,t-1}$ and $\hat{u}_{2,t-1}$ are one-step-ahead forecasting errors for models 1 and 2, respectively.

(DGP2)[4] $$\theta_t = \begin{cases} \theta_0, & t = 1, \ldots, \dfrac{T}{2} \\ \theta_1, & t = \dfrac{T}{2} + 1, \ldots, T \end{cases}$$

Further, the values of α_0 and γ_0 are set as follows:

[4] We have also considered the DGP with two structural breaks as: $\theta_t = \theta_0$, for $t = 1, \ldots, \frac{T}{3}$, $\theta = \theta_1$, for $t = \frac{T}{3} + 1, \ldots, \frac{2T}{3}$, and $\theta = \theta_2$, for $t = \frac{2T}{3} + 1, \ldots, T$. Here $(\alpha_0, \gamma_0) = (0.2, 0.2)$; $(\alpha_1, \gamma_1) = (0.5, 0,5)$; $(\alpha_2, \gamma_2) = (0.8, 0.8)$. The simulation results are similar to DGP2 and are available upon request.

$$\begin{cases}\text{DGP1:} & (\alpha_0, \gamma_0) = (0.3, 0.5) \\ \text{DGP2:} & (\alpha_0, \gamma_0) = (0.2, 0.2); (\alpha_1, \gamma_1) = (0.8, 0.8)\end{cases}$$

In addition, rather than discrete breaks in the DGPs shown above, we consider the following stochastic time variation in the parameters:

$$\textbf{(DGP3)} \qquad \theta_t = \theta_{t-1} + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \sigma_\varepsilon = 0.01$$

where the parameter sequence θ_t is initialized at 0. This kind of data generating process is addressed in Rossi (2003) as well.

To evaluate size, let $\beta = 0$, $\forall t$. To evaluate power, β is set at 0.1. It is clear that for DGP1, we have the same settings as the DGP-I in Clark and McCracken (2001). The number of replications in each experiment is 1000.

SIMULATION RESULTS AND DISCUSSION

Empirical size and power

Table II presents the empirical size results for three different data generating processes, using a nominal size of 5%. We have considered various combinations of *R* (in-sample observations) and *P* (out-of-sample observations). In most settings, both in-sample and out-of-sample tests have reasonable finite-sample size properties with DGP1. It is worth noting that both tests have higher size distortions for DGP2.

We then report the empirical power results in Table III. According to Panel A in Table III, it is clear that all out-of-sample tests have worse power than in-sample tests while DGP1 is used. Moreover, ENC-NEW has the best power over any other out-of-sample tests in all cases. These results concur with Clark and McCracken's (2001) findings. Therefore, according to the simulation results, it seems that the in-sample GC test beats all of the out-of-sample tests, and it may be a more appropriate test to employ when there are no structural changes.

If we consider one breaking point in the data generating process (DGP2), Panel B in Table III shows that there are seven out of nine cases where the ENC-NEW test has higher power than the GC test.[5] In the case of DGP3 (Panel C), however, the results reveal a better performance for the in-sample tests over the out-of-sample tests in most cases. Therefore, it is suggested that the out-of-sample tests are more powerful than the in-sample tests under the appearance of discrete structural breaks, but not for the case of stochastic time variation.

Since we have shown that under discrete structural breaks, the out-of-sample tests may work better than the in-sample tests, a further question may be appealing as well: how does the finding change under different settings of *P*/*R*, the ratio of post-sample portion to in-sample portion? We therefore present the power and size results for different values of *P*/*R* in Table IV. We can see that the power improves as *P*/*R* increases without substantial changes of empirical size.

[5] While accounting for two break points, though not uniformly, we still get that most ENC-NEW tests are more powerful than the GC tests. Moreover, since we focus on a (Granger) causal relationship from *x* to *y*, we simply assume that *y* does not (Granger) cause *x*, i.e., there are no lags of y on the right-hand side. However, including lags of *y* on the right-hand side does not change the results that out-of-sample tests are more powerful than in-sample tests when there exist discrete breaks.

Table II. Empirical size, one-step-ahead forecasts (nominal size = 5%)

(A) DGP1	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.048	0.049	0.054	0.055	0.042	0.045	0.044	0.054	0.039
ENC-NEW	0.053	0.046	0.055	0.058	0.042	0.048	0.057	0.051	0.050
ENC-T	0.057	0.055	0.055	0.060	0.042	0.048	0.055	0.046	0.046
ENC-REG	0.047	0.053	0.052	0.059	0.037	0.046	0.059	0.041	0.049
MSE-F	0.042	0.049	0.049	0.054	0.036	0.043	0.054	0.039	0.050
MSE-T	0.051	0.056	0.050	0.059	0.041	0.044	0.058	0.040	0.044
MSE-REG	0.045	0.053	0.047	0.055	0.042	0.040	0.057	0.043	0.043
CCS	0.041	0.051	0.044	0.042	0.047	0.041	0.048	0.044	0.048
(B) DGP2	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.094	0.110	0.106	0.122	0.101	0.103	0.114	0.102	0.108
ENC-NEW	0.136	0.141	0.129	0.173	0.138	0.133	0.145	0.140	0.173
ENC-T	0.111	0.121	0.117	0.135	0.108	0.115	0.119	0.095	0.125
ENC-REG	0.105	0.120	0.112	0.121	0.100	0.116	0.107	0.092	0.115
MSE-F	0.095	0.104	0.104	0.117	0.093	0.101	0.108	0.091	0.105
MSE-T	0.074	0.089	0.083	0.091	0.086	0.085	0.082	0.060	0.071
MSE-REG	0.061	0.084	0.079	0.084	0.079	0.081	0.073	0.060	0.066
CCS	0.159	0.202	0.189	0.209	0.203	0.215	0.186	0.208	0.228
(C) DGP3	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.061	0.052	0.047	0.055	0.051	0.059	0.045	0.053	0.038
ENC-NEW	0.078	0.049	0.056	0.058	0.047	0.061	0.042	0.052	0.040
ENC-T	0.074	0.063	0.056	0.057	0.048	0.064	0.039	0.050	0.044
ENC-REG	0.073	0.052	0.053	0.054	0.052	0.063	0.036	0.048	0.046
MSE-F	0.068	0.048	0.048	0.054	0.048	0.057	0.040	0.051	0.045
MSE-T	0.076	0.060	0.048	0.068	0.047	0.051	0.034	0.052	0.050
MSE-REG	0.069	0.056	0.048	0.058	0.043	0.048	0.030	0.049	0.048
CCS	0.054	0.059	0.054	0.044	0.044	0.050	0.039	0.052	0.037

Size-adjusted power

In Table III, we have shown that under discrete structural breaks, the out-of-sample ENC-NEW test may work better than the in-sample GC test in terms of empirical (actual) power. However, it is worth noting that the ENC-NEW test sometimes has greater size than the GC test. It is, therefore, of interest to investigate the size-adjusted power under discrete structural breaks (DGP2). Clearly, the results in Table V indicate that the GC test has higher size-adjusted power than out-of-sample tests in most cases.

Although the good performance of out-of-sample tests does not exist in terms of size-adjusted power, this paper still focuses on the empirical (actual) size and power. The reason is as follows. As shown in Clark and West (2004), in order to link the empirical study with the power property of tests, it may be useful to explore the power that is not size-adjusted. This paper tries to shed some light on what may be useful in conducting the empirical strategy. Considering unadjusted power and

Table III. Empirical power, one-step-ahead forecasts (nominal size = 5%)

(A) DGP1	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.190	0.250	0.269	0.366	0.429	0.497	0.520	0.560	0.615
ENC-NEW	0.188	0.244	0.252	0.331	0.410	0.482	0.442	0.516	0.580
ENC-T	0.157	0.222	0.238	0.283	0.375	0.441	0.337	0.422	0.516
ENC-REG	0.139	0.218	0.231	0.278	0.370	0.443	0.331	0.429	0.516
MSE-F	0.138	0.205	0.211	0.264	0.344	0.415	0.350	0.409	0.494
MSE-T	0.124	0.185	0.205	0.230	0.286	0.363	0.242	0.309	0.395
MSE-REG	0.123	0.175	0.204	0.217	0.279	0.349	0.234	0.305	0.392
CCS	0.090	0.163	0.203	0.197	0.272	0.351	0.218	0.281	0.351
(B) DGP2	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.321	0.411	0.455	0.541	0.658	0.686	0.750	0.784	0.821
ENC-NEW	0.374	0.425	0.477	0.594	0.661	0.708	0.718	0.769	0.846
ENC-T	0.310	0.392	0.443	0.530	0.628	0.668	0.585	0.690	0.789
ENC-REG	0.305	0.392	0.432	0.519	0.625	0.668	0.589	0.692	0.785
MSE-F	0.295	0.371	0.404	0.500	0.573	0.629	0.565	0.635	0.756
MSE-T	0.223	0.311	0.355	0.379	0.490	0.555	0.393	0.488	0.612
MSE-REG	0.205	0.306	0.352	0.375	0.484	0.546	0.388	0.474	0.611
CCS	0.357	0.467	0.533	0.585	0.654	0.714	0.622	0.724	0.810
(C) DGP3	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.143	0.218	0.237	0.279	0.368	0.422	0.444	0.477	0.521
ENC-NEW	0.147	0.217	0.241	0.273	0.346	0.408	0.382	0.432	0.483
ENC-T	0.136	0.213	0.224	0.245	0.331	0.395	0.295	0.377	0.430
ENC-REG	0.125	0.205	0.225	0.238	0.328	0.391	0.294	0.369	0.420
MSE-F	0.123	0.190	0.205	0.232	0.304	0.352	0.316	0.362	0.406
MSE-T	0.122	0.164	0.183	0.208	0.258	0.309	0.225	0.283	0.327
MSE-REG	0.106	0.160	0.180	0.200	0.252	0.304	0.215	0.277	0.329
CCS	0.098	0.148	0.178	0.166	0.238	0.292	0.195	0.250	0.294

size may help researchers be aware of the (exact) benefit and cost they may encounter in their empirical study.

A two-stage procedure

According to this simple Monte Carlo simulation study, we may conclude that an out-of-sample test is preferred when there are discrete structural breaks in non-Granger-causal coefficients. In contrast, when the coefficients are stable or the instability is stochastically time-variant, the in-sample test performs better. Thus, it may be intuitive to consider the following two-stage procedure to deal with the causality test:

(1) First test whether there are discrete structural changes.
(2) If we reject the null of no structural breaks, an out-of-sample Granger causality test could be applied; otherwise, an in-sample test may be preferred.

Table IV. Empirical power and size (nominal size = 5%), DGP2

	(1) Empirical power								
	$R = 100$ $P = 20$	40	60	80	100	120	140	160	180
GC	0.394	0.422	0.483	0.518	0.541	0.589	0.631	0.658	0.676
ENC-NEW	0.358	0.411	0.491	0.530	0.594	0.634	0.648	0.661	0.684
ENC-T	0.275	0.320	0.397	0.449	0.530	0.565	0.577	0.628	0.642
ENC-REG	0.266	0.320	0.387	0.426	0.519	0.565	0.575	0.625	0.643
MSE-F	0.304	0.331	0.377	0.411	0.500	0.537	0.535	0.573	0.606
MSE-T	0.227	0.257	0.292	0.302	0.379	0.440	0.442	0.490	0.515
MSE-REG	0.219	0.240	0.273	0.289	0.375	0.437	0.435	0.484	0.514
CCS	0.216	0.323	0.444	0.479	0.585	0.625	0.626	0.654	0.698
	(2) Empirical power								
	$R = 100$ $P = 20$	40	60	80	100	120	140	160	180
GC	0.114	0.095	0.119	0.116	0.122	0.104	0.116	0.101	0.099
ENC-NEW	0.146	0.134	0.146	0.158	0.173	0.146	0.166	0.138	0.130
ENC-T	0.124	0.115	0.108	0.115	0.135	0.108	0.124	0.108	0.100
ENC-REG	0.115	0.103	0.104	0.102	0.121	0.105	0.122	0.100	0.104
MSE-F	0.135	0.111	0.102	0.104	0.117	0.094	0.111	0.093	0.087
MSE-T	0.120	0.100	0.079	0.084	0.091	0.074	0.073	0.086	0.076
MSE-REG	0.113	0.093	0.075	0.069	0.084	0.069	0.068	0.079	0.074
CCS	0.152	0.169	0.185	0.200	0.209	0.207	0.219	0.203	0.203

Table V. Size-adjusted power, one-step-ahead forecasts (size = 5%), DGP2

(A) DGP1	$R = 50$			$R = 100$			$R = 200$		
	$P = 40$	80	100	$P = 100$	160	200	$P = 120$	160	200
GC	0.229	0.286	0.341	0.364	0.549	0.541	0.575	0.633	0.726
ENC-NEW	0.209	0.260	0.286	0.322	0.510	0.523	0.519	0.618	0.654
ENC-T	0.205	0.249	0.272	0.310	0.464	0.523	0.434	0.571	0.656
ENC-REG	0.202	0.254	0.307	0.338	0.502	0.542	0.434	0.539	0.672
MSE-F	0.212	0.263	0.318	0.322	0.469	0.511	0.446	0.536	0.636
MSE-T	0.179	0.214	0.259	0.272	0.394	0.408	0.283	0.450	0.561
MSE-REG	0.183	0.222	0.270	0.269	0.405	0.421	0.287	0.447	0.557
CCS	0.143	0.225	0.266	0.316	0.415	0.450	0.375	0.443	0.524

Ideally, we may expect to see that the above procedure gives the best power since it takes advantage of both in-sample and out-of-sample tests. However, because the tests of structural changes in the first stage have their own power–size property, it is unclear how well the procedure performs in practice. We may further consider the following data generating process:[6]

[6] I would like to thank a referee for suggesting the experiments I conduct here.

Table VI. Empirical power and size (nominal size = 5%), DGP4

	(1) Empirical power								
	$R = 100$ $P = 20$	40	60	80	100	120	140	160	180
GC	0.277	0.309	0.339	0.378	0.403	0.428	0.469	0.475	0.512
OOS	0.239	0.286	0.328	0.364	0.405	0.425	0.457	0.464	0.490
2STAGE	0.258	0.299	0.340	0.379	0.413	0.434	0.477	0.474	0.506
	(2) Empirical power								
	$R = 100$ $P = 20$	40	60	80	100	120	140	160	180
GC	0.068	0.070	0.061	0.064	0.069	0.068	0.070	0.073	0.069
OOS	0.088	0.077	0.082	0.080	0.094	0.091	0.085	0.083	0.070
2STAGE	0.078	0.074	0.079	0.077	0.091	0.089	0.084	0.086	0.073

(DGP4)[7] Generate simulated data from DGP1, DGP2 and DGP3 with probability 1/3, 1/3 and 1/3, respectively

Clearly, in DGP4, we draw from a mixture of distributions in which one is stable (DGP1) and the others are unstable (DGP2 and DGP3). We then compare the power ($\beta = 0.1$ as well) of the proposed two-stage procedure (2STAGE) with naively (1) always using the in-sample test (GC) and (2) always using the out-of-sample ENC-NEW test (OOS). In the first stage, we apply the Andrews and Ploberger (1994) structural break tests with approximate asymptotic p-values proposed by Hansen (1997). Five percent is chosen as the significant level to reject the null of no structural breaks. The results are reported in Table VI.

According to Table VI, it is obvious that the proposed two-stage procedure may outperform the naive in-sample and out-of-sample tests in most cases. And in all cases considered, the two-stage procedure has not been the worst choice. Hence, we may conclude that the two-stage procedure works well.

AN EMPIRICAL EXAMPLE

In this section, we use out-of-sample and in-sample causality tests to determine whether domestic saving is useful in predicting domestic investment. Since Feldstein and Horioka's (1980) seminal paper, it is well known that there is a strong positive correlation between saving and investment rates across countries. Their finding of a close saving–investment link was further confirmed by numerous studies and interpreted as evidence of international capital immobility. Rather than running cross-sectional regressions, we would like to provide time series evidence on the saving–investment relationship by investigating Taiwan's quarterly data.

[7] Alternatively, generating simulated data from DGP1 with probability 1/2 and from DGP2 with probability 1/2 gives similar results.

Table VII. Empirical example: in-sample and out-of-sample tests for Granger causality

	Test statistics	Critical values
In-Sample GC	1.77	3.11 (10%), 4.37 (5%), 9.02 (1%)
ENC-NEW	3.64**	2.35 (10%), 3.28 (5%), 5.52 (1%)
ENC-T	1.71**	1.09 (10%), 1.46 (5%), 2.17 (1%)
ENC-REG	1.48**	1.09 (10%), 1.46 (5%), 2.17 (1%)
MSE-F	1.22*	0.24 (10%), 1.45 (5%), 4.35 (1%)
MSE-T	0.31*	0.06 (10%), 0.37 (5%), 1.00 (1%)
MSE-REG	0.25*	0.06 (10%), 0.37 (5%), 1.00 (1%)
CCS	10.17***	2.71 (10%), 3.84 (5%), 6.63 (1%)

$\log(I)_t = \alpha + \Sigma_{j=1}^{n} \beta_j \log(I)_{t-j} + \Sigma_{j=1}^{n} \gamma_j \log(S)_{t-j} + \varepsilon_t$. Lags($n$) is chosen to be 5 by the Akaike information criterion (AIC). *, ** and *** denote the rejection of the null under 10%, 5% and 1% significance level, respectively. The critical values of out-of-sample tests come from Clark and McCracken (2001) and McCracken (2004).

Our quarterly data, which covers 1961:3 to 2002:2 with a total of 164 observations, is Taiwan's private investments and private net savings. Data comes from the Macroeconomic Statistics Database of the Directorate-General of Budget, Accounting and Statistics, Taiwan. All variables are in logarithms.

First, the results of the augmented Dicky–Fuller tests suggest that both investments and savings reject the hypothesis of unit roots.[8] We then test for structural breaks in individual time series using Andrews and Ploberger's (1994) test. In addition, we also test for common structural breaks following Bai *et al*. (1998). Both tests suggest structural breaks in Taiwan's investments and savings. In particular, the Sup-W statistics, break date and 90% confidence interval of break data in the Bai *et al*. (1998) test are 17.50, 1974:1 and [1971:2, 1976:4], respectively. The common structural break of investments and savings in 1974:1 seems plausible because of the oil crisis in the 1970s.

Since there may exist structural breaks in our data, it is more appropriate to employ out-of-sample tests to determine whether domestic saving Granger causes domestic investment. The in-sample and post-sample portions are set to be $R = 84$ and $P = 80$, which makes $P/R \simeq 1.0$. Moreover, the recursive scheme of estimations is used. In order to make a comparison, we also employ the in-sample GC test. The bivariate VAR model of investment and saving with five lags was chosen by use of the Akaike information criterion (AIC). The details of the empirical model and the results are presented in Table VII. Clearly, the in-sample GC test fails to reject the non-Granger causality while all out-of-sample tests reject the null under either 5% or 10% significance levels. In other words, according to Taiwan's time series data, domestic savings may be helpful in predicting domestic investment. Since we have already shown that under discrete structural breaks, out-of-sample tests may be more powerful than in-sample tests, this example may have shown that a possible Granger causal relationship can be detected by out-of-sample tests while the in-sample GC test fails to reject the null of non-causality.

[8] Results for tests of unit roots and structural breaks are available upon request.

CONCLUDING REMARKS

In this paper, we use a simple Monte Carlo simulation study to investigate a conjecture that the out-of-sample tests may be more powerful than the in-sample tests under the appearance of time-varying parameters. The simulation results indicate that out-of-sample tests may work better than in-sample tests in terms of empirical (actual) power when there are discrete structural breaks in non-Granger-causal coefficients. However, we are also aware that this good performance of out-of-sample tests does not hold in terms of size-adjusted power.

We then propose a two-stage procedure to deal with the causality test. (a) Test first whether there are discrete structural changes. (b) If we reject the null of no structural breaks, out-of-sample Granger causality tests could be applied; otherwise, in-sample tests may be more appropriate. A further Monte Carlo study suggests that relative to naively always using in-sample tests or always using out-of-sample tests, the two-stage procedure works well.

Guided by this two-stage procedure, we investigate the investment–saving relationship in Taiwan by first conducting a test of structural breaks. We found that there may exist structural breaks in the data, thus out-of-sample tests should be applied. The empirical results imply that Taiwan's domestic savings may be helpful in predicting domestic investments and have illustrated that a possible Granger causal relationship may be detected by out-of-sample tests while the in-sample GC test fails to reject the null of non-causality.

ACKNOWLEDGEMENTS

I would like to thank Bruce E. Hansen and Kenneth D. West for valuable discussions and suggestions. I have also benefited from the comments of an anonymous referee and the editor. Any remaining errors are my own responsibility.

REFERENCES

Andrews DWK, Ploberger W. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* **62**(6): 1383–1414.

Ashley R, Granger C, Schmalensee R. 1980. Advertising and aggregate consumption: an analysis of causality. *Econometrica* **48**(5): 1149–1168.

Awad MA, Goodwin BK. 1998. Dynamic linkages among real interest rates in international capital markets. *Journal of International Money and Finance* **17**: 881–907.

Bai J, Lumsdaine RL, Stock JH. 1998. Testing for and dating common breaks in multivariate time series. *Review of Economic Studies* **63**: 395–432.

Chao J, Corradi V, Swanson N. 2001. A out-of-sample test for Granger causality. *Macroeconomic Dynamics* **5**(4): 598–620.

Clark TE, McCracken MW. 2001. Tests of equal forecast accuracy and ecompassing for nested models. *Journal of Econometrics* **105**: 85–110.

Clark TE, McCracken MW. 2005. The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics* **124**: 1–31.

Clark TE, McCracken MW. 2003b. The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence. Working Paper, RWP 03-06, Federal Reserve Bank of Kansas City.

Clark TE, West KD. 2004. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Working Paper, University of Wisconsin-Madison.

Diebold F, Mariano R. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Ericsson NR. 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: an exposition, extensions, and illustration. *Journal of Policy Modeling* **14**(4): 465–495.
Feldstein M, Horioka C. 1980. Domestic saving and international capital flows. *Economic Journal* **90**(358): 314–329.
Granger CW. 1969. Investigating causal relations by econometric model and cross spectral methods. *Econometrica* **37**: 424–438.
Granger CW, Newbold P. 1977. *Forecasting Economic Time Series*. Academic Press: Orlando, FL.
Hansen BE. 1997. Approximate asymptotic *P* values for structural-change tests. *Journal of Business and Economic Statistics* **15**(1): 60–67.
Harvey D, Leybourne S, Newbold P. 1998. Tessts for forecast encompassing. *Journal of Business and Economic Statistics* **16**: 254–259.
Inoue A, Kilian L. 2003. In-sample or out-of-sample tests of predictability: which one should we use? Working Paper, University of Michigan.
Lettau M, Ludvigson S. 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* **LVI**(3): 815–849.
McCracken M. 2004. Asymptotics for out-of-sample tests of causality. Working Paper, Department of Economics, University of Missouri-Columbia.
Rapach DE, Weber CE. 2004. Financial variables and the simulated out-of-sample forecastability of U.S. output growth since 1985: an encompassing approach. *Economic Inquiry* **42**(4): 717–738.
Rapach DE, Weber CE. 2004. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. Working Paper, Department of Economics, Saint Louis University.
Rossi B. 2003. Optimal tests for nested model selection with underlying parameter instability. Working paper, Duke University.

Author's biography:
Shiu-Sheng Chen is an assistant professor in the Department of Economics, National Taiwan University. He holds a PhD in Economics from the University of Wisconsin-Madison. His research interests are macroeconomics, international finance, time series modelling and forecasting.

Author's address:
Shiu-Sheng Chen, Department of Economics, National Taiwan University, No. 21 Hsu-Chow Rd, 100 Taipei, Taiwan.

