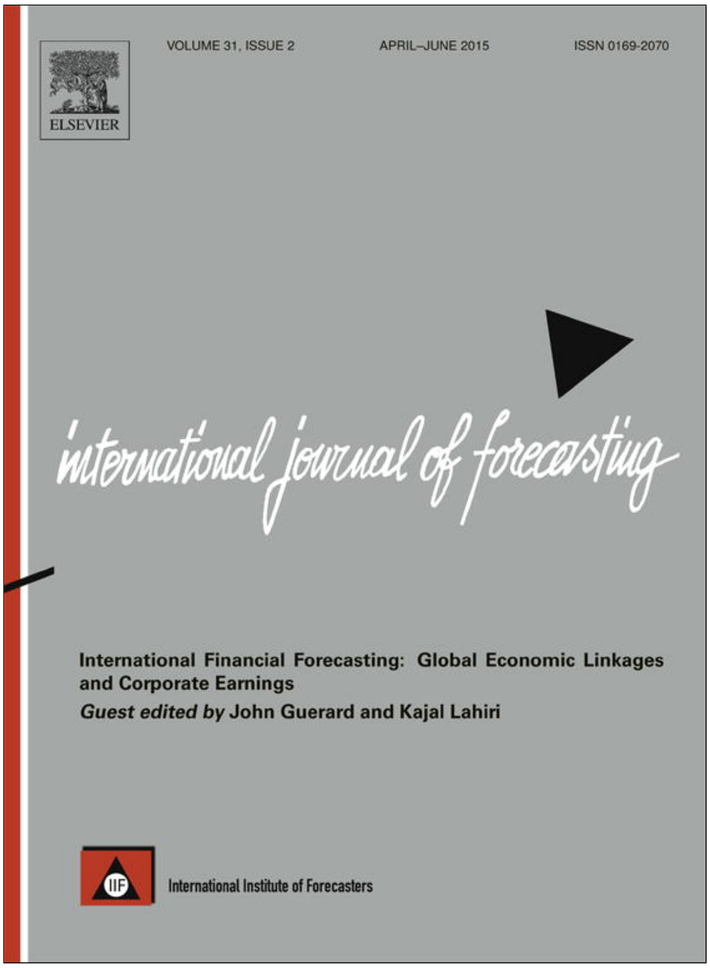


VOLUME 31, ISSUE 2 APRIL–JUNE 2015 ISSN 0169-2070

ELSEVIER

international journal of forecasting

International Financial Forecasting: Global Economic Linkages and Corporate Earnings

Guest edited by John Guerard and Kajal Lahiri

International Institute of Forecasters

International Journal of Forecasting 31 (2015) 488–500

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Comparing the effectiveness of traditional vs. mechanized identification methods in post-sample forecasting for a macroeconomic Granger causality analysis

Haichun Ye[a,*,1], Richard Ashley[b], John Guerard[c]

[a] School of Economics, Shanghai University of Finance and Economics, Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, 111 Wuchuan Road, Shanghai 200434, China

[b] Department of Economics (0316), Virginia Tech, Blacksburg, VA 24061, United States

[c] McKinley Capital Management, LLC, 3301 C Street, Suite 500, Anchorage, AK 99503, United States

ARTICLE INFO

Keywords:
Post-sample forecasting
Post-sample Granger causality
Identification methods

ABSTRACT

We identify forecasting models using both a traditional, partially judgmental method and the mechanized Autometrics method. We then compare the effectiveness of these two different identification methods for post-sample forecasting, in the context of a relatively large-scale exemplar of macroeconomic post-sample Granger causality testing. This example examines the Granger causal relationships among four macroeconomically important endogenous variables – monthly measures of aggregate income, consumption, consumer prices, and the unemployment rate – embedded in a six-dimensional information set which also includes two interest rates, both of which are taken to be weakly exogenous in this context. We find that models indentified by the traditional method tend to have better post-sample forecasting abilities than analogous models identified using the mechanized method, and that the analysis done using the traditional identification method generates stronger evidence for post-sample Granger causality among the four endogenous variables.

1. Introduction

In-sample Granger causality analysis is typically based on an F-test of the null hypothesis that the coefficients on the putatively-causing variates in a particular VAR model equation are all zero. It has long been known that such tests are so routinely misleading as to be of doubtful usefulness. As was discussed by Racine and Parmeter (2013, Section 1) and Efron (1982, Chapter 7), this is an inevitable consequence of the fact that these in-sample F tests are inherently based on model fitting errors. These fitting errors – the magnitudes of which are, by definition, being minimized by the estimation process itself – correspond to what Efron calls 'apparent' rather than 'true' errors. Consequently, a comparison of the post-sample forecasting effectiveness over varying information sets has long been the methodology of choice in this area, albeit implemented in a variety of ways: see Ashley (2003), Ashley, Granger, and Schmalensee (1980), Guerard (1985), and Thomakos and Guerard (2004). The reader is referred to Ashley and Tsang (2014) and Ashley and Ye (2012) for a review of this literature.[2]

[*] Corresponding author.
E-mail address: haichunye@gmail.com (H. Ye).

[1] Haichun Ye appreciates the financial support of the Shanghai Pujiang Program.

[2] Notably, these papers discuss recent criticisms of the post-sample forecasting testing framework, including the developing realization that

http://dx.doi.org/10.1016/j.ijforecast.2014.08.004
0169-2070/

While it is well-known that a key step in post-sample forecasting is to identify relevant time series models over both the full and restricted information sets, very little is known about the effectiveness of different model identification methods in post-sample forecasting. In this study, we address this issue by identifying models in two interestingly distinct ways and then comparing the effectiveness of the two model identification approaches. Specifically, as per Ashley and Ye (2012), the models (over both the full and restricted information sets) are first identified in the somewhat *ad hoc* "large-to-small" manner commonly identified with David Hendry: one starts with as complicated a model as the data set will support (i.e., a vector autoregression in each included variable, utilizing all lags out to at least the seasonal lag), then pares down this formulation by eliminating statistically insignificant terms, starting with the largest, least plausible, lags.[3] It is common (and sensible) to use some judgment in this process, so we will identify this below as the "partially judgmental" identification procedure. For example, an isolated statistically significant lag structure term at lag twelve is likely to be worth retaining in a model for monthly data, whereas such a term at a lag of eight or eleven is not.[4] Alternatively, analogous models (over both the full and restricted information sets) are also identified and estimated using the "Autometrics" mechanized model specification procedure introduced by Doornik and Hendry (2007) and currently implemented in the *Oxmetrics* software program. Both of these model identification algorithms – along with their sample fits to the data considered here – are described at greater length in Section 2 below. The relative effectiveness of these two identification algorithms in post-sample forecasting is then examined in Section 3, in the context of a new, relatively large-scale exemplar of Granger causality testing.

Ashley and Ye (2012) test for post-sample Granger causality between the median growth rate in these 31 sub-components of the US Consumer Price Index (i.e., the monthly CPI inflation rate) and the inter-quartile range of these 31 sub-components (i.e., the monthly dispersion in the inflation rates across the 31 categories), but this is only a bivariate analysis. Here we employ six, arguably more broadly interesting, US macroeconomic aggregates:

- Aggregate real income
  This variable is defined as the monthly growth rate of seasonally adjusted real disposable personal income, and is denoted "y_t" below.
- Aggregate real household consumption spending
  This variable is defined as the monthly growth rate of seasonally adjusted real personal consumption expenditures, and is denoted "c_t" below.
- CPI inflation rate
  This variable is defined as the monthly growth rate of the seasonally unadjusted consumer price index (CPI), and is denoted "π_t" below.
- Civilian unemployment rate
  This variable is defined as the monthly change in the seasonally unadjusted civilian unemployment rate, and is denoted "Δun_t" below.

These time series are taken to be endogenous, which is to say, potentially Granger-caused by each other and/or by the final two time series considered; lags in these last two time series are therefore taken to be weakly exogenous:[5]

- Short-term interest rate
  This variable is defined as the monthly change in the seasonally unadjusted 3-month Treasury bill rate, and is denoted "$\Delta tbill_t$" below.
- Long-term interest rate[6]
  This variable is defined as the monthly change in the seasonally unadjusted yield on 10-year Treasury bonds, and is denoted "$\Delta tbond_t$" below.

These data are all used in un-deseasonalized form whereever possible (i.e., for $\pi_{t,}$ Δun_t, $\Delta tbill_t$, and $\Delta tbond_t$), as the Bureau of Economic Analysis' de-seasonalization method employs a two-sided filter which distorts causal inferences.

The data sources, summary statistics, time plots, and sample correlograms for these six time series are presented in Tables 1 and 2 and Fig. 1. The changes in Δun_t, $\Delta tbill_t$, and $\Delta tbond_t$ are used instead of their levels because these levels data are so highly persistent that a unit root in the levels time series cannot be rejected credibly on standard tests. The null hypothesis of a unit root is rejected at the 1% level for all six time series (as defined above) using both the *ADF* and *PP* tests; see Table 3.[7]

Consequently, we proceeded on the assumption that all six time series, as formulated above, are $I(0)$.

In this setting, we find that models identified by the "partially judgmental" data procedure tend not to fit the sample data as well, but produce smaller post-sample mean squared forecast errors (MSFE) than those identified by the Autometrics algorithm. The analysis based on the traditional, partially judgmental model specification approach yields stronger evidence for post-sample

particular care must be taken (as is done below) in choosing a statistical test for post-sample forecasting improvements in the context of nested models. Another problem with post-sample testing is the *ad hoc* nature of the data split between a model identification/estimation sub-period and a post-sample model evaluation sub-period. Ashley and Tsang (2014) and Racine and Parmeter (2013) have each developed model validation methods based on cross-validation which surmount this obstacle, for modest sample lengths and large sample lengths, respectively; a follow-on paper to the present work will apply the Racine–Parmeter cross-validation model validation procedure to the (large-sample) data set and models examined here.

[3] If reasonably feasible, it is a good idea to exceed the seasonal lag at the outset, as a multiplicatively seasonal model can be expected to yield terms beyond the seasonal lag when one identifies an additive model.

[4] See Ashley (2012, Section 14.4) for a discursive example.

[5] We are by no means asserting that fluctuations in the other four variables do not Granger- cause fluctuations in these two interest rates, we are simply not testing for these causal links.

[6] The yields used here as $tbill_t$ and $tbond_t$ are taken from the St. Louis Federal Reserve website as the secondary market rate for a three-month Treasury bill and the constant maturity rate for a ten-year Treasury bond. Measuring yields on such securities is a non-trivial endeavor, with the realized yields being likely to be slightly superior to those used here.

[7] The absence of a strong negative sample autocorrelation at lag one in the correlograms for Δun_t, $\Delta tbill_t$, and $\Delta tbond_t$ confirms that they are not over-differenced. An *ARFIMA* model for the levels variables was not considered, for the reasons given, at length, by Ashley and Patterson (2010).

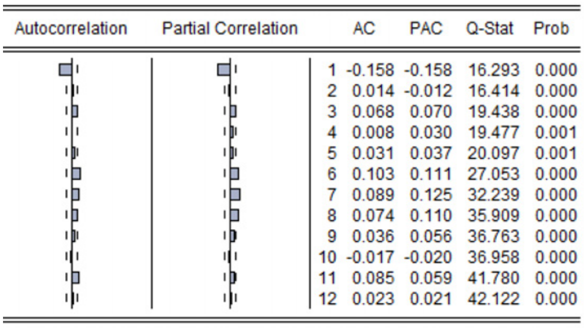
Table 1
Data source and summary statistics.

	c_t	y_t	π_t	Δun_t	$\Delta tbill_t$	$\Delta tbond_t$
Mean	0.269	0.263	0.320	0.002	−0.004	−0.003
Median	0.269	0.257	0.296	0	0.01	0
Maximum	2.382	5.735	1.790	0.9	2.61	1.61
Minimum	−2.764	−5.359	−1.934	−0.7	−4.62	−1.76
Std. Dev.	0.542	0.759	0.356	0.182	0.445	0.286
Skewness	−0.279	−0.119	−0.006	0.499	−1.760	−0.436
Kurtosis	5.781	19.371	6.396	4.737	28.979	8.985

All six monthly series used in this study are retrieved from the FRED II dataset. This table provides summary statistics for the six variables over the full sample period (1959M2–2013M5).

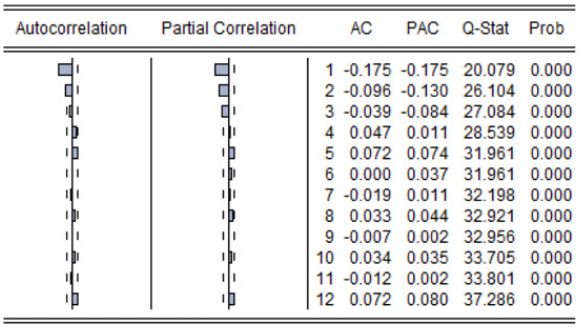
Table 2
Sample correlograms: 1959M1–2013M5.

c_t



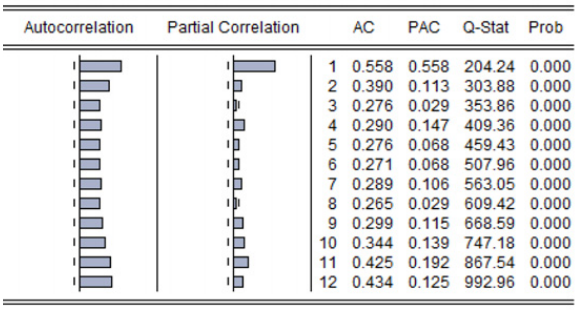
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.158	-0.158	16.293	0.000
		2	0.014	-0.012	16.414	0.000
		3	0.068	0.070	19.438	0.000
		4	0.008	0.030	19.477	0.001
		5	0.031	0.037	20.097	0.001
		6	0.103	0.111	27.053	0.000
		7	0.089	0.125	32.239	0.000
		8	0.074	0.110	35.909	0.000
		9	0.036	0.056	36.763	0.000
		10	-0.017	-0.020	36.958	0.000
		11	0.085	0.059	41.780	0.000
		12	0.023	0.021	42.122	0.000

y_t



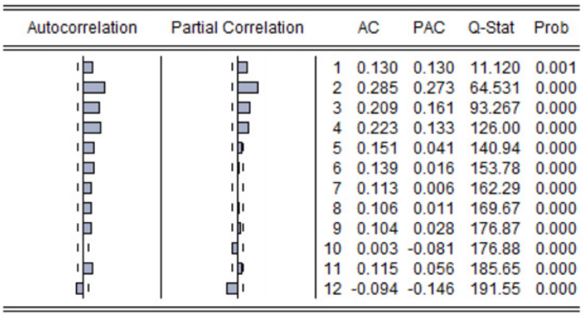
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.175	-0.175	20.079	0.000
		2	-0.096	-0.130	26.104	0.000
		3	-0.039	-0.084	27.084	0.000
		4	0.047	0.011	28.539	0.000
		5	0.072	0.074	31.961	0.000
		6	0.000	0.037	31.961	0.000
		7	-0.019	0.011	32.198	0.000
		8	0.033	0.044	32.921	0.000
		9	-0.007	0.002	32.956	0.000
		10	0.034	0.035	33.705	0.000
		11	-0.012	0.002	33.801	0.000
		12	0.072	0.080	37.286	0.000

π_t



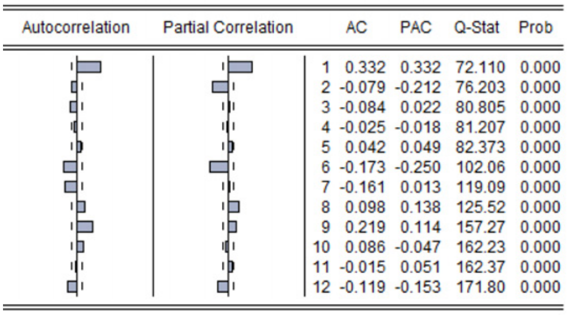
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.558	0.558	204.24	0.000
		2	0.390	0.113	303.88	0.000
		3	0.276	0.029	353.86	0.000
		4	0.290	0.147	409.36	0.000
		5	0.276	0.068	459.43	0.000
		6	0.271	0.068	507.96	0.000
		7	0.289	0.106	563.05	0.000
		8	0.265	0.029	609.42	0.000
		9	0.299	0.115	668.59	0.000
		10	0.344	0.139	747.18	0.000
		11	0.425	0.192	867.54	0.000
		12	0.434	0.125	992.96	0.000

Δun_t



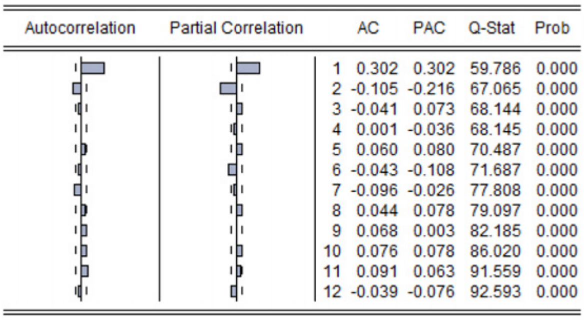
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.130	0.130	11.120	0.001
		2	0.285	0.273	64.531	0.000
		3	0.209	0.161	93.267	0.000
		4	0.223	0.133	126.00	0.000
		5	0.151	0.041	140.94	0.000
		6	0.139	0.016	153.78	0.000
		7	0.113	0.006	162.29	0.000
		8	0.106	0.011	169.67	0.000
		9	0.104	0.028	176.87	0.000
		10	0.003	-0.081	176.88	0.000
		11	0.115	0.056	185.65	0.000
		12	-0.094	-0.146	191.55	0.000

$\Delta tbill_t$



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.332	0.332	72.110	0.000
		2	-0.079	-0.212	76.203	0.000
		3	-0.084	0.022	80.805	0.000
		4	-0.025	-0.018	81.207	0.000
		5	0.042	0.049	82.373	0.000
		6	-0.173	-0.250	102.06	0.000
		7	-0.161	0.013	119.09	0.000
		8	0.098	0.138	125.52	0.000
		9	0.219	0.114	157.27	0.000
		10	0.086	-0.047	162.23	0.000
		11	-0.015	0.051	162.37	0.000
		12	-0.119	-0.153	171.80	0.000

$\Delta tbond_t$



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.302	0.302	59.786	0.000
		2	-0.105	-0.216	67.065	0.000
		3	-0.041	0.073	68.144	0.000
		4	0.001	-0.036	68.145	0.000
		5	0.060	0.080	70.487	0.000
		6	-0.043	-0.108	71.687	0.000
		7	-0.096	-0.026	77.808	0.000
		8	0.044	0.078	79.097	0.000
		9	0.068	0.003	82.185	0.000
		10	0.076	0.078	86.020	0.000
		11	0.091	0.063	91.559	0.000
		12	-0.039	-0.076	92.593	0.000

Granger causality among the variables considered in this study. We believe that the differences in the results from post-sample Granger causality tests are a consequence of the mechanically-produced model specifications being less able to forecast post-sample.

The plan of the remainder of this paper is as follows. The models identified and estimated using these two approaches are described and compared in Section 2. The forecasting results for the full information set, based on the two model identification approaches, are compared in

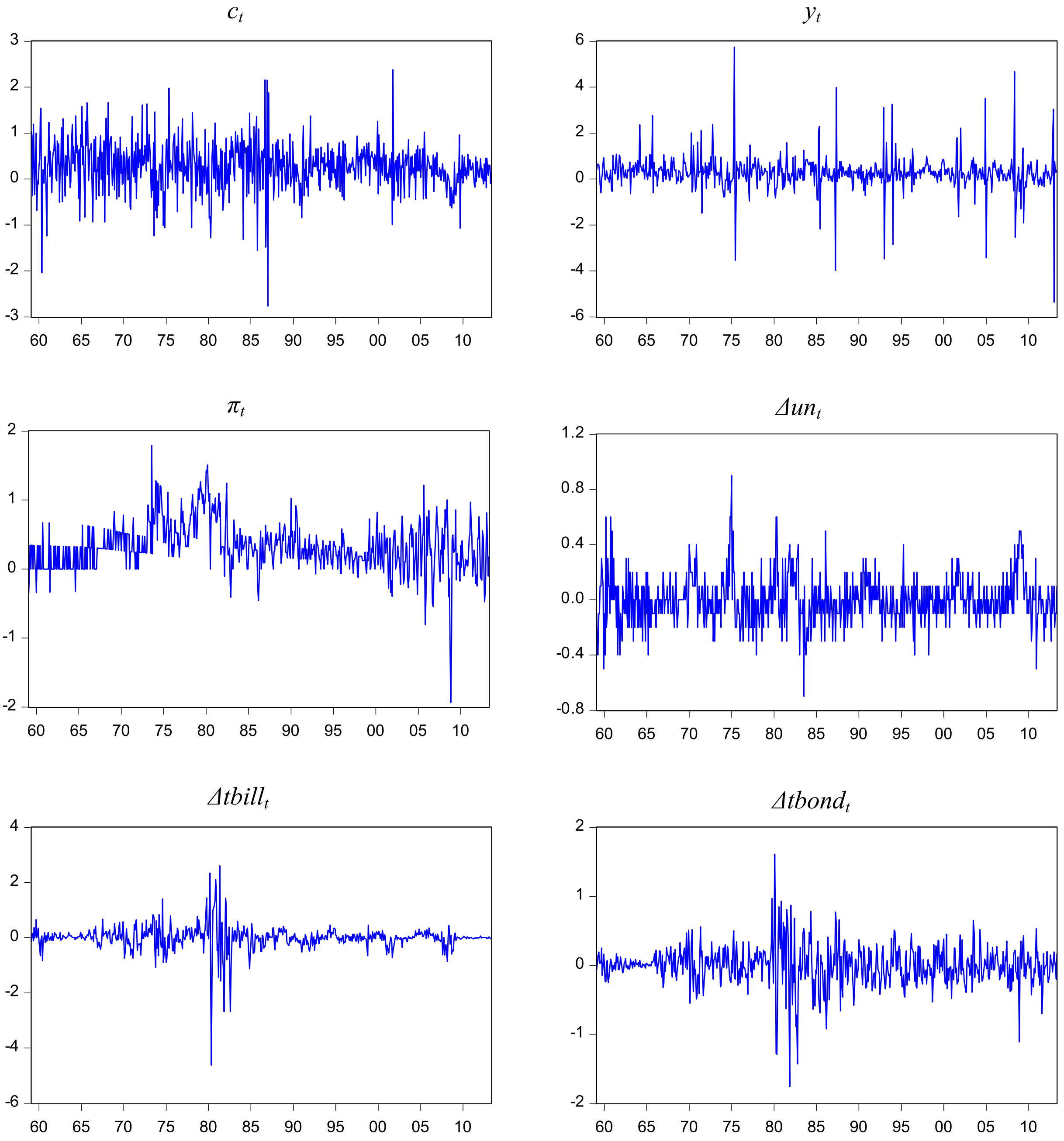


Fig. 1. Data time plots: 1959M1–2013M5.

Section 3; and the post-sample Granger causality testing and results are described in Section 4. Section 5 concludes the paper with overall comments on the causal relationships found and on the relative effectiveness of the two model identification procedures employed.

2. Model identification and estimation

This section describes the two alternative model identification procedures and presents their respective in-sample model coefficient estimates.

Prior to model identification and estimation, we reserved the first 12 observations (February 1959 to January 1960) for creating lagged variables. We then used the 395 sample observations from February 1960 to December 1992 for model identification/estimation, and reserved the remaining 245 observations, over the period from January 1993 to May 2013, solely for post-sample forecasting and Granger-causality tests, although model coefficient estimate updating is allowed (and done) throughout this post-sample forecasting period.[8]

[8] When using the Autometrics approach to identify model specifications, the first 24 observations are used to create lags and the in-sample estimations are conducted over the period 1961M2–1992M12, with a total of 383 observations. This particular sample vs. post-sample split decision was made here at the outset in order to obtain a reasonably representative post-sample testing period which is also sufficiently lengthy to allow the post-sample MSE reduction tests to have adequate power. As was noted above, a companion paper using the present data is in preparation, in which this sample-splitting decision is side-stepped, using the cross-validation methods described by Ashley and Tsang (2014) (for

Table 3
Unit root test results.

	c_t	y_t	π_t	Δun_t	$\Delta tbill_t$	$\Delta tbond_t$
ADF test	-5.663^{***}	-11.588^{***}	-2.599^{***}	-6.844^{***}	-5.982^{***}	-7.2124^{***}
PP test	-29.697^{***}	-31.175^{***}	-15.905^{***}	-26.110^{***}	-17.240^{***}	-18.120^{***}

Notes: These results utilize the full data set, 1959M2–2013M5. The *AIC* is used to select the lag length in the *ADF* test; these tests assume that an intercept is included in the test equation for each time series.
*** Indicates significance at the 1% level.

To carry out the Granger causality tests between two variables, we compare an unrestricted model, which includes lags in the putatively "causing" variable as explanatory variables, to a restricted model, from which these lags are excluded. For example, when testing for Granger-causality from consumption (c_t) to income (y_t), we simply compare the unrestricted model of income, in which lags of consumption are included as explanatory variables, to the restricted model of income, in which the consumption lags are not used in the model identification process. In both restricted and unrestricted models we also control for the other (possibly causative) variables, and for short-term and long-term interest rates when these additional variables have been identified as belonging in the model for income.

Here, we use two different approaches to identify the unrestricted model for each of the four endogenous variables. We first identify the models in the "large-to-small" manner commonly identified with David Hendry. This identification procedure is referred to below as "partially judgmental", and consists of the following steps:

(1) for each endogenous variable, start with an equation which includes 12 lags of its own, 12 lags in each of the five remaining variables, and outlier dummies whenever a plot of the fitting errors indicates that some of these are necessary;

(2) one at a time, remove all of the statistically insignificant lag 12 terms (including the 12th lag in the dependent variable) in alphabetical (or inverse-alphabetical) order;

(3) next, remove all of the non-significant lag 11 terms one at a time in the same way, including those that are significant *per se* but not part of a coherent lag structure (a "coherent lag structure" including a term at lag 11 would probably also have statistically significant terms at lags 10, 9, 8, etc.);

(4) repeat Step (3) for lag 10, and so on;

(5) remove any outlier dummies that have become statistically insignificant.

As a final step, diagnostic checks (such as plotting the fitting errors) should also be applied.[9] Two of the co-authors independently applied this "partially judgmental" identification algorithm to the four endogenous variables (y_t, c_t, π_t and Δun_t), and obtained essentially identical model specifications, which are given as:[10]

$$\begin{aligned} y_t &= \alpha + \sum_{i=1}^{3}\beta_i y_{t-i} + \sum_{i=1}^{3}\delta_i c_{t-i} + \sum_{i=1}^{2}\phi_i \pi_{t-i} \\ &\quad + \sum_{i=1}^{3}\Delta tbond_{t-i} + D75M5 + D87M4 + \varepsilon_t \end{aligned}$$

$$\begin{aligned} c_t &= \varphi + \sum_{i=1}^{8}\gamma_i c_{t-i} + \kappa y_{t-1} + \sum_{i=1}^{2}\lambda_i \Delta un_{t-i} \\ &\quad + \sum_{i=1}^{2}\varpi_i \pi_{t-i} + \eta_t \end{aligned}$$

$$\begin{aligned} \pi_t &= \chi + \sum_{i=1}^{4}\theta_i \pi_{it} + \theta_{12}\pi_{t-12} + \sum_{i=1}^{2}\vartheta_i y_{t-i} \\ &\quad + \sum_{i=1}^{4}\rho_i c_{t-i} + \sum_{i=1}^{2}\sigma_i \Delta tbill_{t-i} + \varsigma\Delta tbond_{t-1} \\ &\quad + D73M8 + \nu_t \end{aligned}$$

$$\begin{aligned} \Delta un_t &= \mu + \sum_{i=1}^{4}\tau_i \Delta un_{t-i} + \tau_{12}\Delta un_{t-12} \\ &\quad + \sum_{i=1}^{3}\psi_i c_{t-i} + \psi_{12} c_{t-12} + \xi_t. \end{aligned}$$

The three "restricted information set" models were obtained similarly for each of the four dependent variables, dropping one of the other three potentially causative explanatory variables – out of y_t, c_t, π_t and Δun_t – from consideration in each case. The coefficient estimates, standard error estimates and I, the usual best-practice measure of sample fit, adjusted for model complexity, are all listed in Table 4a for each of the four unrestricted models.

Using just the data until December 1992, as was also the case for the previous "partially judgmental" model identifications, the remaining co-author then identified models for each of these four endogenous time series (over both the full and restricted information sets) using the "Autometrics" mechanized model specification procedure introduced by Doornik and Hendry (2007) and currently implemented in the *Oxmetrics* software described by Castle and Shepard (2009), Doornik and Hendry (2009a,b) and Hendry (2000).

modest sample lengths) and Racine and Parmeter (2013) (for long sample lengths).

[9] Such plots would warn of outliers or grotesque heteroscedasticity, although the latter is less important because of the use of robust standard error estimates. In general, the inclusion of a sufficient number of lagged dependent and explanatory variables eliminates serial correlation in the errors.

[10] $D75M5_t$, $D87M4_t$ and $D73M8_t$ are outlier dummies for the three months of May 1975, April 1987 and August 1973, respectively. Where variables at the seasonal lag (12) were found to be significant, we then also considered terms at lags 13 and 14, as such terms could arise from a multiplicative seasonal model.

Table 4a
Model coefficient estimates using the partially-judgmental identification procedure.

	Dependent variable			
	y_t	c_t	π_t	Δun_t
y_{t-1}	-0.259^{**} (0.111)	0.130^{**} (0.052)	-0.025 (0.016)	
y_{t-2}	-0.231^{***} (0.071)		0.054^{***} (0.019)	
y_{t-3}	-0.106^{**} (0.048)			
c_{t-1}	0.050 (0.050)	-0.293^{***} (0.068)	0.037^{*} (0.021)	-0.076^{***} (0.014)
c_{t-2}	0.104^{**} (0.045)	-0.143^{**} (0.063)	0.002 (0.021)	-0.062^{***} (0.015)
c_{t-3}	0.132^{**} (0.057)	-0.063 (0.066)	-0.034 (0.021)	-0.037^{**} (0.016)
c_{t-4}		-0.097^{*} (0.055)	0.061^{***} (0.021)	
c_{t-5}		-0.021 (0.045)		
c_{t-6}		0.061 (0.050)		
c_{t-7}		0.118^{**} (0.054)		
c_{t-8}		0.115^{**} (0.053)		
c_{t-12}				-0.033^{***} (0.012)
π_{t-1}	-0.373^{***} (0.112)	-0.161 (0.110)	0.269^{***} (0.060)	
π_{t-2}	-0.197^{*} (0.100)	-0.349^{***} (0.103)	0.202^{***} (0.051)	
π_{t-3}			-0.031 (0.049)	
π_{t-4}			0.196^{***} (0.049)	
π_{t-12}			0.209^{***} (0.043)	
Δun_{t-1}		-0.055 (0.189)		-0.094^{*} (0.052)
Δun_{t-2}		-0.399^{**} (0.199)		0.162^{***} (0.056)
Δun_{t-3}				0.133^{***} (0.049)
Δun_{t-4}				0.195^{***} (0.048)
Δun_{t-12}				-0.171^{***} (0.045)
$\Delta tbill_{t-1}$			-0.046 (0.037)	
$\Delta tbill_{t-2}$			0.088^{***} (0.033)	
$\Delta tbond_{t-1}$	0.208^{*} (0.120)		0.192^{***} (0.055)	
$\Delta tbond_{t-2}$	0.038 (0.109)			
$\Delta tbond_{t-3}$	0.242^{**} (0.099)			
$D73M8_t$			1.310^{***} (0.050)	
$D75M5_t$	5.642^{***} (0.166)			
$D87M4_t$	-4.013^{***} (0.181)			
BIC	729.9494	761.1385	35.0789	−230.3391

Notes: All models are estimated using the in-sample period 1960M2 to 1992M12. Constant terms are included but are not reported. $D73M8_t$, $D75M5_t$ and $D87M4_t$ are month dummies. Robust standard errors are reported in parentheses.
[*] Indicates significance at the 10% level.
[**] Indicates significance at the 5% level.
[***] Indicates significance at the 1% level.

Autometrics, as described by Doornik (2009), is the third generation of the Hendry (2000) *GETS* ("general-to-specific") model selection procedure, which has gradually evolved into the Autometrics algorithm over the past 20–30 years. The Autometrics algorithm has several primary ingredients: (1) the general unrestricted model, GUM, is the starting point for all analysis; (2) multiple path searches are performed; (3) the encompassing test is performed; (4) diagnostics checks are employed; and (5) a tiebreaker procedure is employed. Since the estimated GUM is checked by diagnostics tests, the GUM is statistically well-behaved. The k insignificant variables identified by the algorithm create k paths for model reduction, beginning with the variables with the lowest absolute t-values. The encompassing test is used to ensure that the current model encompasses the GUM, while other diagnostic tests are used to examine issues of normality, residual correlation, and residual ARCH. An automated "tiebreaker" routine is then used to allow this fully automated procedure to decide on a final model specification.[11]

The starting point for the initial model in Autometrics is the entire space generated by using all variables in the regression model. The most statistically insignificant variable, on the basis of the absolute t-value, is eliminated before estimating the next model. The subnodes are then reordered, with the most insignificant variable first. The search algorithms in Autometrics: (1) prune the model, removing one variable at each reduction; (2) bunch several statistically insignificant variables together; and (3) chop the least statistically significant variables from the branches of the model. The final ("terminal") model cannot be reduced on the basis of the criteria adopted. The regression tree analysis is ordered uniquely, and one can determine the minimal branch that can be deleted in order to produce a different model. Diagnostic checking is used only after the terminal model has been reached.

In Autometrics, the initial GUM is estimated. Dummy variables are then added to deal with possible outliers, with the regressors being tested at a large significance

[11] Model coefficient estimates are updated in subsequent ("recursive") post-sample forecasting – as are those of the models obtained using the partially judgmental method described earlier – but the model *specifications* are not updated using either approach.

level; if the null hypothesis that they enter with a coefficient of zero is not rejected, then diagnostic tests are performed. The starting point for the current model is the GUM. If all of the variables are statistically significant, then the algorithm pauses and the diagnostic testing is updated. In an ideal world, the regressors in the GUM should pass all diagnostic tests. If this is not the case, then the p-value is raised for each failed diagnostic test statistic. Terminal candidates are collected as the search procedures run, and previously identified sub-trees are skipped. Terminal candidates that fail either diagnostic tests or the encompassing test are removed.

Table 4b reports the in-sample estimates of the unrestricted models for the four endogenous variables identified using Autometrics procedures, again including a *BIC* value for each estimated model.

In Table 5 we provide a condensed summary comparing the in-sample model identification/estimation results provided by the partially-judgmental vs. Autometrics model identification algorithms. Broadly speaking, while the two approaches usually (but not always) agree on the variables to be included in each equation, they differ with respect to the lag length of each variable, whether they control for changes in short-term/long-term interest rates, and also in the outlier dummy variables included.

On the other hand, it is worth noting that the model specification algorithm choice is not entirely of no consequence with regard to Granger-causality among the variables. In particular, the partially judgmental specifications include a lagged y_t in the equations for c_t, whereas the Autometrics specifications do not, and the Autometrics specifications include lagged values of Δun_t in the y_t and π_t equations, whereas the partially judgmental specifications do not. Thus, if one uses the partially judgmental model identification algorithm, then the possibility of finding Granger causality running from Δun_t to either y_t or π_t is eliminated at the outset, whereas the use of the Autometrics algorithm eliminates at the outset the (Keynesian) possibility of Granger causality running from y_t to c_t. Of course, this result does not eliminate the possibility that lagged values of one or more of the other variables may be "proxying" for a lagged y_t, nor the possibility that this Keynesian-type causal link is operating primarily on a contemporaneous (within a month) basis.

Based on the observed *BIC* values, the Autometrics model specifications are generally distinctly preferable, in terms of their fit to the sample data.[12] On the other hand, precisely as one might expect, the partially judgmental model specifications seem more intuitively plausible to us than the corresponding Autometrics-based specifications. For example, the Autometrics-chosen unrestricted model

Table 4b
Model coefficient estimates using the Doornik–Hendry "Autometrics" identification procedure.

	Dependent variable			
	y_t	c_t	π_t	Δun_t
y_{t-1}	-0.139^{*} (0.077)		-0.052^{***} (0.018)	
y_{t-2}	-0.168^{***} (0.047)			
y_{t-4}			-0.050^{***} (0.013)	
y_{t-13}	-0.080^{**} (0.038)			
y_{t-21}	-0.119^{***} (0.038)			
y_{t-24}	0.054 (0.038)			
c_{t-1}		-0.230^{***} (0.052)	0.055^{***} (0.020)	-0.062^{***} (0.014)
c_{t-2}		-0.198^{***} (0.047)		-0.049^{***} (0.015)
c_{t-3}		-0.114^{**} (0.049)		-0.033^{**} (0.013)
c_{t-4}	-0.088^{*} (0.053)		0.937^{***} (0.020)	
c_{t-6}	-0.133^{***} (0.046)			
c_{t-7}	-0.071 (0.049)			
c_{t-8}	0.043 (0.044)	0.157^{***} (0.044)		
c_{t-11}	0.140^{***} (0.039)			
c_{t-12}	0.112^{***} (0.039)			
c_{t-13}	0.076 (0.047)			
c_{t-16}	0.118^{**} (0.048)			-0.042^{***} (0.012)
c_{t-17}	0.059 (0.040)			
c_{t-19}				-0.062^{***} (0.011)
c_{t-20}	0.062 (0.040)	0.106^{***} (0.040)		
c_{t-22}			0.059^{***} (0.016)	
c_{t-23}	-0.067^{*} (0.040)			
c_{t-24}		-0.086^{**} (0.040)		
π_{t-1}	-0.222^{***} (0.082)		0.194^{***} (0.052)	
π_{t-2}		-0.407^{***} (0.097)	0.124^{***} (0.047)	
π_{t-4}		-0.262^{***} (0.090)	0.119^{***} (0.043)	
π_{t-7}			0.116^{***} (0.041)	
π_{t-9}			0.186^{***} (0.045)	

(continued on next page)

[12] The *BIC* value is calculated as: $BIC = -2\ln(L)+k\ln(N)$, where $\ln(L)$ is the maximized log-likelihood of the model, k is the number of parameters estimated and N is the number of observations. To ensure that the *BIC* values are comparable between two model identification methods, we also re-estimated the partially judgmental model specifications over the sample period 1961M2–1992M12 and obtained the *BIC* values for the income, consumption, inflation and unemployment rate equations as 710.7858, 723.2877, 34.3288 and −245.1629, respectively. The Autometrics method still yields smaller *BIC* values than the partially judgmental method.

Table 4b (*continued*)

	Dependent variable			
	y_t	c_t	π_t	Δun_t
π_{t-10}	-0.264^{***} (0.092)	-0.217^{**} (0.100)		
π_{t-12}	-0.238^{**} (0.108)		0.210^{***} (0.039)	
π_{t-14}		0.388^{***} (0.105)		
π_{t-15}			0.130^{***} (0.038)	
π_{t-16}	0.180^{*} (0.098)			
π_{t-18}			-0.157^{***} (0.039)	
π_{t-24}	0.105 (0.085)			
Δun_{t-2}	-0.493^{***} (0.141)	-0.434^{***} (0.155)		
Δun_{t-3}				0.111^{**} (0.045)
Δun_{t-4}			-0.174^{***} (0.056)	0.205^{***} (0.044)
Δun_{t-5}			-0.179^{***} (0.057)	
Δun_{t-6}	-0.160 (0.159)			
Δun_{t-7}		0.432^{***} (0.162)		-0.105^{**} (0.044)
Δun_{t-11}			-0.215^{***} (0.056)	
Δun_{t-12}				-0.200^{***} (0.043)
Δun_{t-13}	0.334^{**} (0.134)			
Δun_{t-15}	0.262 (0.164)			
$\Delta tbill_{t-2}$			0.067^{**} (0.030)	-0.050^{**} (0.020)
$\Delta tbill_{t-3}$				0.036^{*} (0.021)
$\Delta tbill_{t-7}$	-0.153^{**} (0.072)			
$\Delta tbill_{t-11}$	-0.130^{**} (0.056)			
$\Delta tbill_{t-15}$	0.085^{*} (0.048)			
$\Delta tbill_{t-16}$				0.043^{***} (0.014)
$\Delta tbill_{t-19}$	0.134^{***} (0.047)			
$\Delta tbill_{t-21}$		-0.104^{***} (0.039)	0.066^{***} (0.022)	
$\Delta tbill_{t-23}$			0.044^{**} (0.022)	0.033^{*} (0.017)
$\Delta tbill_{t-24}$	0.137^{**} (0.065)			
$\Delta tbond_{t-1}$			0.161^{***} (0.035)	
$\Delta tbond_{t-3}$	0.161^{*} (0.089)			
$\Delta tbond_{t-6}$	0.131 (0.094)		-0.091^{**} (0.039)	
$\Delta tbond_{t-7}$	0.236^{**} (0.114)			
$\Delta tbond_{t-11}$	0.112 (0.111)			
$\Delta tbond_{t-12}$	0.202^{**} (0.103)			
$\Delta tbond_{t-14}$				0.093^{***} (0.029)
$\Delta tbond_{t-15}$	-0.272^{**} (0.106)			
$\Delta tbond_{t-18}$			0.083^{**} (0.035)	
$\Delta tbond_{t-20}$	-0.252^{***} (0.091)			0.101^{***} (0.025)
$\Delta tbond_{t-21}$	0.272^{***} (0.086)			
$\Delta tbond_{t-24}$	-0.274^{**} (0.114)			
$D65M9_t$	2.381^{***} (0.119)			
$D65M10_t$		1.562^{***} (0.133)		
$D66M5_t$		-1.367^{***} (0.094)		
$D68M3_t$		1.689^{***} (0.088)		
$D70M4_t$	1.756^{***} (0.105)			
$D72M10_t$	2.391^{***} (0.132)			
$D73M8_t$			1.421^{***} (0.040)	
$D75M1_t$				0.800^{***} (0.042)
$D75M5_t$	3.431^{***} (0.664)			
$D75M6_t - D75M5_t$	-2.518^{***} (0.510)			
$D80M4_t$		-1.685^{***} (0.124)		
$D85M10_t$		-1.582^{***} (0.108)		

(*continued on next page*)

for c_t includes isolated (albeit statistically significant) terms in c_{t-8} and Δun_{t-7}, which we find a bit unappealing.

Clearly, these issues need to be addressed via a consideration of the post-sample forecasting performances of the models, to which we now turn.

Table 4b (*continued*)

	Dependent variable			
	y_t	c_t	π_t	Δun_t
$D87M1_t$		-3.422^{***} (0.175)		
$D87M4_t$	-4.571^{***} (0.216)			
$D92M12_t$	2.333^{***} (0.126)			
BIC	680.76	621.1008	−16.8867	−287.1568

Notes: All models are estimated using the in-sample period 1961M2 to 1992M12. Constant terms are included in all models except for the π_t regression. $D1965M9_t$, $D65M10_t$, $D66M5_t$, $D68M3_t$, $D70M4_t$, $D72M10_t$, $D73M8_t$, $D75M1_t$, $D75M5_t$, $D75M6_t$, $D80M4_t$, $D85M10_t$, $D87M1_t$, $D87M4_t$, and $D92M12_t$ are month dummies. Robust standard errors are reported in parentheses.

* Indicates significance at the 10% level.

** Indicates significance at the 5% level.

*** Indicates significance at the 1% level.

Table 5

Condensed comparison of model specifications.

	Partially judgmental	Autometrics
y_t equation:		
Lagged c_t	√	√
Lagged π_t	√	√
Lagged Δun_t		√
c_t equation:		
Lagged y_t	√	
Lagged π_t	√	√
Lagged Δun_t	√	√
π_t equation:		
Lagged c_t	√	√
Lagged y_t	√	√
Lagged Δun_t		√
Δun_t equation:		
Lagged c_t	√	√
Lagged π_t		
Lagged y_t		

Notes: Intercepts and lagged dependent variables are included in all models.

3. Post-sample forecasting

Based on the two model specifications identified above, we next obtained one-step-ahead post-sample forecasts from the restricted and unrestricted models for each of the four endogenous variables, using a rolling scheme with a fixed forecasting window of width equal to the number of in-sample observations.[13]

More explicitly, for each of the partially judgmental specifications, the model parameters are first estimated on the sample running from 1960M2 to 1992M12 and used to produce a forecast for each endogenous variable at date 1993M1, then re-estimated on the sample running from 1960M3 to 1993M1 and used to produce forecasts at date 1993M2, and so forth. (The Autometrics-based forecasting was almost identical, except that the initial window began twelve months later.) The corresponding (rolling) one-step-ahead forecast errors were then used to compute the post-sample mean squared forecast error (MSFE) for each of the four endogenous variables, using both the unrestricted and restricted models for that variable.

We also constructed naïve benchmark forecasts (intercept-only models, corresponding to a constant growth rate or change) for each of the four endogenous variables and then compared the post-sample MSFEs from these naïve forecasting models to those from both the restricted and unrestricted models.

In addition to these forecasting results over the entire post-sample period (i.e., from 1993M1 to 2013M5), we also computed post-sample MSFE results for two subsets of this period: a "pre-crisis" period (1993M1–2007M12) and a "crisis-plus-aftermath" period (2008M1–2013M5).

These results, with separate columns for the two model identification methods, are all reported in Tables 6a–6d. The naïve forecast MSFE values are displayed in the top row of each table, and the MSFE results for both the restricted and unrestricted models are presented in the immediately following rows as the ratios to the results for naïve forecasting models.

Regardless of which model identification approach is used, we find that the restricted and unrestricted models are able to produce more accurate forecasts than the naïve model in most cases, and that the forecasts for the crisis-plus-aftermath period (2008M1–2013M5) are generally less accurate than those for the pre-crisis period (1993M1–2007M12).

Notably, the post-sample MSFE results from the models based on the Autometrics specification algorithm are always larger than those from the partially judgmental model specification approach. While it is not clear that these differences are statistically significant, the uniformity of the results strongly suggests that the "informed common sense" utilized in the partially judgmental model specification method yields better models, in terms of their post-sample forecasting ability, than does the current state-of-the-art in mechanical model specification methodology.

Some specific post-sample forecasting results are worth elaborating on. For the income equation, including lagged consumption generally reduces the MSFE, while including the inflation rate or changes in the unemployment rate actually increases the MSFE somewhat. For the consumption equation, including lagged values of the inflation rate leads to a rise in the post-sample MSFE, while including income (in the case of the partially judgmental models) increases the MSFE of the consumption forecasts over the pre-crisis period, though it does reduce it slightly in the crisis-plus-aftermath period. In the case of the Autometrics models, including the change in the unemployment rate raises the MSFE of the consumption forecasts over both the entire post-sample period and the pre-crisis period; this variable decreases the MSFE by about 3% in the crisis-plus-aftermath period.

For the inflation rate equation, including lagged values of consumption or changes in the unemployment rate tends to increase the MSFE overall. In contrast, including lagged income reduces the MSFE over the entire post-sample period in the model identified by the partially

[13] This window comprised 395 observations for forecasts using the partially judgmental specifications and 383 observations for the Autometrics-based forecasts (because the latter considered variables lagged 24 months rather than just 12).

Table 6a
Model forecasting results for y_t (post-sample *MSFE* ratio vs. naïve model).

	Post-sample period					
	1993M1–2013M5		1993M1–2007M12		2008M1–2013M5	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Naïve model	0.730		0.503		1.358	
Full information set	0.892	0.989	0.875	1.017	0.910	0.961
Omitting lagged c_t	0.950	0.991	0.918	1.007	0.982	0.974
Omitting lagged π_t	0.885	0.968	0.862	1.008	0.909	0.926
Omitting lagged Δun_t	–	0.967	–	0.975	–	0.959

Notes: "Naïve model" entries are rolling window one-step-ahead post-sample *MSFE* values for the naïve model; the other results are all displayed as a ratio to the corresponding naïve model *MSFE*. The column heading "P. Judg." stands for "partially judgmental".

Table 6b
Model forecasting results for c_t (post-sample *MSFE* ratio vs. naïve model).

	Post-sample period					
	1993M1–2013M5		1993M1–2007M12		2008M1–2013M5	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Naïve model	0.142		0.141			
Full information set	1.057	1.195	0.993	1.115	1.231	1.413
Omitting lagged y_t	1.025	–	0.947	–	1.237	–
Omitting lagged π_t	0.942	1.010	0.948	0.968	0.927	1.124
Omitting lagged Δun_t	1.047	1.150	0.983	1.040	1.224	1.450

Notes: "Naïve model" entries are rolling window one-step-ahead post-sample *MSFE* values for the naïve model; the other results are all displayed as a ratio to the corresponding naïve model *MSFE*. The column heading "P. Judg." stands for "partially judgmental".

Table 6c
Model forecasting results for π_t (post-sample *MSFE* ratio vs. naïve model).

	Post-sample period					
	1993M1–2013M5		1993M1–2007M12		2008M1–2013M5	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Naïve model	0.151226		0.113339		0.256145	
Full information set	0.651	0.703	0.677	0.743	0.618	0.655
Omitting lagged y_t	0.655	0.683	0.673	0.715	0.634	0.644
Omitting lagged c_t	0.642	0.694	0.666	0.720	0.612	0.661
Omitting lagged Δun_t	–	0.689	–	0.721	–	0.649

Notes: "Naïve model" entries are rolling window one-step-ahead post-sample *MSFE* values for the naïve model; the other results are all displayed as a ratio to the corresponding naïve model *MSFE*. The column heading "P. Judg." stands for "partially judgmental".

Table 6d
Model forecasting results for Δun_t (post-sample *MSFE* ratio vs. naïve model).

	Post-sample period					
	1993M1–2013M5		1993M1–2007M12		2008M1–2013M5	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Naïve model	0.024		0.017		0.045	
Full information set	0.836	0.916	1.006	1.097	0.657	0.726
Omitting lagged c_t	0.875	0.936	1.018	1.020	0.724	0.847
Omitting lagged π_t	–	–	–	–	–	–
Omitting lagged Δun_t	–	–	–	–	–	–

Notes: "Naïve model" entries are rolling window one-step-ahead post-sample *MSFE* values for the naïve model; the other results are all displayed as a ratio to the corresponding naïve model *MSFE*. The column heading "P. Judg." stands for "partially judgmental".

judgmental approach, although it raises the MSFE in the model identified by the Autometrics approach. While both model specifications imply that including the lagged income increases the MSFE over the pre-crisis period (by about 0.6% in the partially judgmental specification and 4% in the Autometrics specification), the two identification approaches differ with regard to the forecasting power of lagged income for inflation over the crisis-plus-aftermath period: including lagged income reduces the post-sample MSFE by about 2.5% in the partially judgmental specification but raises it by 1.7% in the Autometrics specification.

Finally, with regard to the equation for the change in the unemployment rate, including consumption reduces the post-sample MSFE for forecasting Δun_t over the entire post-sample period by 4.5% in the partially judgmental specification and 2% in the Autometrics specification. While both model specifications imply that including lagged consumption reduces the MSFE during the crisis-plus-aftermath period, the forecasting results for the pre-crisis period are different: including lagged consumption reduces the MSFE in the partially judgmental specification but raises it in the Autometrics specification.

Because, in general, they are able to forecast better post-sample, the partially judgmental specification results seem to have a clearer interpretation than those based on the Autometrics specifications. However, before framing the differential forecasting results over differing information sets explicitly in terms of Granger causality, it is appropriate to test whether the forecasting improvements found are statistically significant; this is the topic of the next section.

4. Post-sample Granger causality testing

Based on the above post-sample forecasting results, we now proceed to the post-sample statistical testing for Granger causality among the four endogenous variables. Specifically, in each case, we examine whether the post-sample MSFE from the unrestricted model for a particular endogenous variable is smaller than that obtained from a restricted model which omits the past values of the putatively causative variable; this done by testing the null hypothesis that these two MSFE values are equal.

For example, to test for Granger-causality from consumption (c_t) to income (y_t), we compare the MSFE for the unrestricted model of income to that for the restricted model that omits lagged values of consumption. If the former is smaller than the latter and the null hypothesis of equality can be rejected, then one can conclude that consumption has predictive power for income. Such a result is then taken to be evidence for Granger causality from consumption to income.[14]

As per the theoretical results of McCracken (2007), when the restricted and unrestricted models are nested, the asymptotic distributions of the Granger-Newbold and Diebold–Mariano test statistics are significantly non-normal, and hence, can lead to serious testing size distortions. To eliminate this problem, we use McCracken's F-type test statistic:

$$\text{MSE} - F = P \sum (e_{r,t}^2 - e_{u,t}^2) \Big/ \sum e_{u,t}^2,$$

where $e_{r,t}$ and $e_{u,t}$ are the post-sample forecast errors from the restricted and unrestricted models, respectively, and P is the number of post-sample observations. As was shown by Clark and McCracken (2001) and McCracken (2007), this test is also more powerful than the Diebold and Mariano test when the models are nested.

As McCracken (2007) pointed out, the asymptotic distribution of the *MSE-F* test statistic itself is non-standard and depends on the forecasting scheme (fixed, rolling or recursive), the number of excess parameters in the nesting model, and the ratio of the number of out-of-sample observations to the number of in-sample observations. Here, as per Ashley and Ye (2012), we sidestep these problems by using Monte Carlo simulations to compute p-values for rejecting the null hypothesis of equal out-of-sample forecasting effectiveness for the restricted and unrestricted models. Simulated data for each of the four endogenous variables are generated by bootstrap re-sampling from the fitting errors of the unrestricted models for each of these variables. In view of the probable presence of heteroskedasticity in the data, this re-sampling was done using the 'wild' bootstrap proposed by Goncalves and Kilian (2004). Specifically, denoting the fitting errors from the unrestricted models for income, consumption, the inflation rate and the change in unemployment rate as τ_t, υ_t, η_t and ω_t, respectively, we draw a sequence of i.i.d. innovations ε_t, $t = 1, 2, \ldots, T$, from the standard normal distribution, and use $\varepsilon_t\ \tau_t$, $\varepsilon_t \upsilon_t$, $\varepsilon_t \eta_t$ and $\varepsilon_t \omega_t$ as the bootstrapped innovations to generate an artificial data set of 652 observations.[15] The restricted and unrestricted models are then re-estimated and the *MSE-F* test statistic is calculated for the new data set. This completes one bootstrap replication. A total of 5000 such replications are done, and the p-value for the *MSE-F* test statistic is computed as the proportion of the generated test statistic values which exceed the test statistic value obtained by using the actual sample data to estimate models and produce the post-sample forecasts.

Tables 7a–7c report the *MSE-F* test statistic values and the null hypothesis rejection p-values for the entire post-sample period, the pre-crisis subsample and the crisis-cum-aftermath subsample, respectively. Based on forecasting throughout the entire post-sample period and using the post-sample forecasts based on the partially judgmental model specifications, there is evidence of Granger causality running from consumption growth rates to income growth rates, from income growth rates to the inflation rate, and also from consumption growth rates to changes in the unemployment rate. The analogous post-sample forecasts based on the Autometrics model specifications only yield evidence for consumption growth rates Granger-causing changes in unemployment over this period. Turning to the pre-crisis subset of this period, the partially judgmental specifications still find Granger causality from consumption growth rates to income growth rates and from consumption growth rates to changes in the unemployment rate, whereas the Autometrics specifications yield no evidence of Granger causality among these four variables at all. In the crisis-cum-aftermath subset of the post-sample period, both the partially judgmental and Autometrics specifications yield evidence of Granger causality from consumption growth rates to income growth rates and from consumption growth rates to changes in unemployment. Over this latter subset of the post sample period, the partially judgmental model specifications yield evidence that income growth rates Granger-cause inflation, but only at the 10% significance level; the models based on the Autometrics specifications yield no evidence for this causality link at all.

5. Conclusions

This paper investigates the impacts of different model identification methods on post-sample forecasting. In particular, we identify forecasting models using two

[14] The MSFE-reduction testing methodology used here is essentially identical to that of Ashley and Ye (2012), which the reader should consult for a more detailed discussion than is given below. In fact, the only differences here are that a noticeably larger number of (substantially more macroeconomically interesting) economic time series are considered in both the unrestricted and restricted models, and that two different model identification schemes are employed and compared.

[15] For simplicity, we fix the values of initial observations at their actual sample values.

Table 7a
Post-sample Granger causality test result summary (using the full post-sample period, 1993M1–2013M5).

	Granger-caused variable							
	y_t		c_t		π_t		Δun_t	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Lagged y_t	–	–	−7.438 (0.975)	–	1.691* (0.069)	−7.052 (0.997)	–	–
Lagged c_t	15.849*** (0.000)	0.509 (0.116)	–	–	−3.340 (0.808)	−3.373 (0.957)	11.270*** (0.000)	5.284** (0.021)
Lagged π_t	−1.822 (0.558)	−5.319 (0.752)	−26.621 (0.998)	−37.910 (1.000)	–	–	–	–
Lagged Δun_t	–	−5.457 (0.988)	−2.269 (0.524)	−9.208 (0.954)	–	−5.039 (0.991)	–	–

Notes: McCracken's *MSE-F* test statistics are reported, with their bootstrapped *p*-values in parentheses. The column heading "P. Judg." stands for "partially judgmental".

* Indicates that the null hypothesis of no Granger causality can be rejected at the 10% significance level.

** Indicates that the null hypothesis of no Granger causality can be rejected at the 5% significance level.

*** Indicates that the null hypothesis of no Granger causality can be rejected at the 1% significance level.

Table 7b
Post-sample Granger causality test result summary (using the pre-crisis post-sample period 1993M1–2007M12).

	Granger-caused variable							
	y_t		c_t		π_t		Δun_t	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Lagged y_t	–	–	−8.347 (0.987)	–	−1.230 (0.627)	−6.708 (0.999)	– –	– –
Lagged c_t	8.989*** (0.000)	−1.635 (0.434)			−2.957 (0.890)	−5.496 (0.997)	2.001** (0.041)	−12.605 (0.997)
Lagged π_t	−2.510 (0.804)	−1.543 (0.335)	−8.249 (0.945)	−23.734 (0.998)	–	–	–	–
Lagged Δun_t	–	−7.344 (0.999)	−1.949 (0.673)	−12.111 (0.998)	–	−5.178 (0.997)	–	–

Notes: McCracken's *MSE-F* test statistics are reported, with their bootstrapped *p*-values in parentheses. The column heading "P. Judg." stands for "partially judgmental".

* Indicates that the null hypothesis of no Granger causality can be rejected at the 10% significance level.

** Indicates that the null hypothesis of no Granger causality can be rejected at the 5% significance level.

*** Indicates that the null hypothesis of no Granger causality can be rejected at the 1% significance level.

Table 7c
Post-sample Granger causality test result summary (using crisis-cum-aftermath post-sample period 2008M1–2013M5).

	Granger-caused variable							
	y_t		c_t		π_t		Δun_t	
	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics	P. Judg.	Autometrics
Lagged y_t	–	–	0.320 (0.142)	–	1.647* (0.055)	−1.015 (0.736)	–	–
Lagged c_t	5.151*** (0.000)	0.923* (0.091)	–	–	−0.642 (0.544)	0.619 (0.119)	6.649*** (0.002)	10.882*** (0.000)
Lagged π_t	−0.066 (0.372)	−2.339 (0.809)	−16.063 (0.999)	−13.264 (0.996)	–	–	–	–
Lagged Δun_t	–	−0.139 (0.343)	−0.377 (0.376)	1.720 (0.113)	–	−0.596 (0.557)	–	–

Notes: McCracken's *MSE-F* test statistics are reported, with their bootstrapped *p*-values in parentheses. The column heading "P. Judg." stands for "partially judgmental".

* Indicates that the null hypothesis of no Granger causality can be rejected at the 10% significance level.

** Indicates that the null hypothesis of no Granger causality can be rejected at the 5% significance level.

*** Indicates that the null hypothesis of no Granger causality can be rejected at the 1% significance level.

different approaches: using a traditional, partially judgmental method, and using the mechanized Autometrics method. We then compare their effectiveness in the specific context of the post-sample forecasting used to complete a relatively large scale macroeconomic Granger causality analysis.

We find that the post-sample forecasting ability of the models identified by the traditional method is generally superior to that of the models identified by the mechanized method. In terms of specific Granger causality testing results, the traditional, partially judgmental model identification method yields statistically significant post-sample evidence for Granger causality running from consumption to income, from income to the inflation rate, and from consumption to changes in the unemployment rate. In contrast, a completely analogous analysis using forecasting models identified using the mechanized Autometrics method only finds weak evidence, at most, for consumption Granger-causing changes in unemployment; this difference in the set of Granger causality results is a consequence of the mechanically-produced model specifications (over both the unrestricted and restricted information sets) being less able to forecast well post-sample.

Overall, we find that the model identification method choice does indeed have a notable impact on both the post-sample forecasting and Granger-causality testing results. In particular – for better or for worse – a bit of experienced human judgment still yields better forecasting models than does the best currently-available mechanical method, at least for this particular data set.

References

Ashley, R. A. (2003). Statistically significant forecasting improvements: how much out-of-sample data is likely necessary? *International Journal of Forecasting*, *19*, 229–240.

Ashley, R. A. (2012). *Fundamentals of applied econometrics*. Hoboken, New Jersey: Wiley.

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

Ashley, R., & Patterson, D. M. (2010). Apparent long memory in time series as an artifact of a time-varying mean: considering alternatives to the fractionally integrated model. *Macroeconomic Dynamics*, *14*, 59–87.

Ashley, R., & Tsang, K. P. (2014). *Credible Granger-causality inference with modest sample lengths: a cross-sample validation approach*. Unpublished manuscript. Available at: http://ashleymac.econ.vt.edu/working_papers/Ashley_Tsang_Cross_Sample_Validation_Granger_Causality.pdf.

Ashley, R., & Ye, H. (2012). On the Granger causality between median inflation and price dispersion. *Applied Economics*, *44*, 4221–4238.

Castle, J., & Shepard, N. (2009). *The methodology and practice of econometrics*. Oxford: Oxford University Press.

Clark, T., & McCracken, M. (2001). Test of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110.

Doornik, J. A. (2009). Autometrics. In J. Castle, & N. Shepard (Eds.), *The methodology and practice of econometrics*. Oxford: Oxford University Press.

Doornik, J. A., & Hendry, D. F. (2007). *PCGIVE*. London: Timberlake Consultants, Ltd..

Doornik, J. A., & Hendry, D. F. (2009a). *Empirical econometric modelling*. Timberlake Consultants, Ltd..

Doornik, J. A., & Hendry, D. F. (2009b). *Modelling dynamic systems*. Timberlake Consultants, Ltd..

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Goncalves, S., & Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, *123*(1), 89–120.

Guerard, J. B., Jr. (1985). Mergers, stock prices, and industrial production: an empirical test of the Nelson hypothesis. In O. D. Anderson (Ed.), *Time series analysis: theory and practice 7* (pp. 239–248). Amsterdam: Elsevier.

Hendry, D. F. (2000). *Econometrics: alchemy or science?* Oxford: Oxford University Press.

McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, *140*(2), 719–752.

Racine, J. S., & Parmeter, C. (2013). Data-driven model evaluation: a test for revealed performance. In A. Ullah, J. S. Racine, & L. Su (Eds.), *Handbook of applied nonparametric and semiparametric econometrics and statistics*. Oxford University Press, URL: http://www.ncsu.edu/cenrep/workshops/documents/modeval.pdf.

Thomakos, D., & Guerard, J. (2004). Naïve, ARIMA, transfer function, and VAR models: a comparison of forecasting performance. *International Journal of Forecasting*, *20*, 53–67.

Haichun Ye earned her Ph.D. in Economics from the University of Oklahoma in 2007.

Richard Ashley earned his Ph.D. in Economics from the University of California, San Diego in 1976. He has published both a monograph, *A nonlinear time series workshop: a toolkit for detecting and identifying nonlinear time series dependence* (Springer, 2000, with Douglas M. Patterson) and an undergraduate/Masters-level textbook, *Fundamentals of applied econometrics* (Wiley, 2012). Richard serves as an Associate Editor of the journal *Econometrics*.

John Guerard, Jr., is Director of Quantitative Research at McKinley Capital Management, in Anchorage, Alaska. He earned his Ph.D. in Finance from the University of Texas, Austin. Mr. Guerard has published several monographs, including *Quantitative Corporate Finance* (Springer, 2007, with Eli Schwartz) and *The handbook of portfolio construction: contemporary applications of Markowitz techniques* (Springer, 2010). John serves an Associate Editor of the *Journal of Investing* and the *International Journal of Forecasting*.