CAN OUT-OF-SAMPLE FORECAST COMPARISONS HELP PREVENT OVERFITTING?

Todd E. Clark

DECEMBER 2000

RWP 00-05

Research Division
Federal Reserve Bank of Kansas City

Todd E. Clark is an assistant vice president and economist at the Federal Reserve Bank of Kansas City. He gratefully acknowledges the helpful comments of Mike McCracken and seminar participants at the Federal Reserve Bank of Kansas City. The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Clark e-mail: todd.e.clark@kc.frb.org

Abstract

This paper shows that out-of-sample forecast comparisons can help prevent data mining-induced overfitting. The basic results are drawn from simulations of a simple Monte Carlo design and a real data-based design similar to those in Lovell (1983) and Hoover and Perez (1999). In each simulation, a general-to-specific procedure is used to arrive at a model. If the selected specification includes any of the candidate explanatory variables, forecasts from the model are compared to forecasts from a benchmark model that is nested within the selected model. In particular, the competing forecasts are tested for equal MSE and encompassing. The simulations indicate most of the post-sample tests are roughly correctly sized, as long as just the in-sample portion of the data are used in model selection. Moreover, the tests have relatively good power, although some are consistently more powerful than others. The paper concludes with an application, modeling quarterly U.S. inflation.

JEL Nos.: C52, C53, E37

Keywords: forecasts, overfitting, model selection, causality

1. Introduction

It is widely recognized that empirical modeling is prone to overfitting. In particular, various forms of data mining may lead a researcher to falsely conclude that some variable x has explanatory power for another variable y.[1] As discussed by Lovell (1983) and Hoover and Perez (1999), the data mining may take the form of a search across candidate models for y. For example, a researcher might search across 10 different x variables to find the one that has the most explanatory power for y. The data mining may also more generally reflect the results of a profession–wide search that has affected the set of candidate variables, a possibility noted by West (1996) and considered in some detail by Denton (1985).

In the hope of reducing the probability of overfitting, many researchers examine out–of–sample forecasts for evidence of predictive power. In the simplest case, if in–sample evidence suggests some x has explanatory power for y, a researcher may construct competing forecasts of y, using one model of y that includes x and another that does not. If x truly has explanatory power for y, forecasts from the model including x should be superior. Accordingly, Ashley, Granger, and Schmalensee (1980) advocate using out–of–sample forecast comparisons to test Granger causality. In practice, using forecast comparisons to determine whether one variable has explanatory power for another has been common since at least the influential work of Meese and Rogoff (1983, 1988).

Although out–of–sample forecast comparisons are widely used, little is known about their effectiveness in preventing overfitting. The extant research on forecasts from nested models has generally focused on a framework in which two pre–specified models are simply compared.[2] In this setting, McCracken (1999) and Clark and McCracken (2000) derive the asymptotic distributions of some basic tests of equal forecast accuracy and forecast encompassing, respectively. Monte

[1] While the term "data mining" in this paper simply refers to searching across candidate models contained within some general specification, Hand (1999) stresses that there are many other forms of data mining.

[2] Although there is now a large literature on the asymptotic and finite–sample properties of equal accuracy and encompassing tests, much of it is focused on non–nested models. The forecasts considered in this paper, however, are from nested models. As noted by McCracken (1999) and Clark and McCracken (2000), for most standard tests whether the models are nested significantly affects the asymptotic distribution.

Carlo simulations in Clark and McCracken show that, in an environment without data mining, the tests considered have good size and power properties. In the same setting, Chao, Corradi, and Swanson (2000) develop an encompassing–type out–of–sample test of causality that has a standard distribution, and show the test also has reasonable finite–sample size and power properties. The results of McCracken (2000) suggest the effectiveness of post–sample tests in these previous studies may carry over to environments with data mining. McCracken establishes conditions under which some simple forms of out–of–sample inference are free from data mining–induced biases, and then shows that, with data mining, an out–of–sample ARCH test has better finite–sample size and power properties than the standard in–sample test.[3]

Accordingly, this paper examines the effectiveness of out–of–sample forecast comparisons in preventing data mining–induced overfitting in finite samples. The basic results are drawn from simulations of a simple Monte Carlo design and a real data–based design similar to those in Lovell (1983) and Hoover and Perez (1999). In each simulation draw, a simple general–to–specific modeling procedure is used to arrive at a model. If that selected specification includes any of the candidate explanatory variables x, forecasts from the model are compared to forecasts from a benchmark model that is restricted to exclude all x variables. In particular, the forecasts are tested for equal MSE and encompassing. This paper presents simulation evidence on the frequency with which, in finite samples, the tests correctly indicate the selected x variables have no predictive power (evidence on "size"), as well as on the frequency with which the tests correctly determine that some x variables have explanatory power (evidence on "power").

The simulation results show that out–of–sample forecast comparisons can help avoid overfitting. Most of the post–sample tests — which are only conducted if the model selection procedure indicates some x variables have explanatory power — are roughly correctly sized in both of the

[3] McCracken's (2000) asymptotic results do not cover the complicated problem posed by the out–of–sample forecast comparisons considered in this paper.

simulation designs considered in this paper. To be useful, though, the forecast comparisons must be made in the way recommended by Ashley, Granger, and Schmalensee (1980). In particular, the available data should be divided into in–sample and out–of–sample portions, and just the in–sample portion should be used in the specification search. The approach used by many researchers, in which all of the data are first used in some form of model determination and then an in–sample portion is used to reestimate the chosen models and generate out–of–sample forecasts, is of relatively little help in preventing overfitting.[4] Finally, the simulations in this paper show that the powers of the post–sample forecast tests follow a simple ranking. Overall, power appears to be relatively good.

To provide further evidence on the effectiveness of out–of–sample forecast comparisons in the presence of data mining, the paper concludes with an application — developing a simple model for quarterly U.S. inflation. A general–to–specific search procedure yields a model relating consumer price inflation to lags of inflation, energy and import price inflation, and the rate of capacity utilization. Tests of equal forecast accuracy and encompassing based on out–of–sample projections for 1990:Q1 through 1999:Q4 indicate energy and import price inflation and capacity utilization have significant predictive power.

While this study focuses on the out–of–sample forecast comparisons that have become commonplace, there are of course other potentially effective strategies for avoiding data mining–induced overfitting.[5] One related strategy, developed by White (2000), is to form post–sample forecast statistics for all models under consideration (not just the best in–sample model and a benchmark model) and then bootstrap p-values. Another strategy is to try to adjust the critical values used in model evaluation. Suppose, for example, that a researcher intends to search across N different x variables to find the best model for y as a function of a single x, and wants the test to have size

[4] Recent examples of studies using all data to select a model and then examining out–of–sample forecasts for a portion of the dataset include Amano and van Norden (1995), Chinn and Meese (1995), Evans and Lyons (1999), and Lettau and Ludvigson (2000).

[5] An additional strategy, advocated by Ericsson and Campos (1999), is to examine recursive t-statistics for evidence of a breakdown in any relationship.

of α percent. Lovell (1983) suggests a reasonable t-statistic to use in gauging the significance of any single x variable is the standard normal critical value for α/N percent.[6] Yet another strategy, advocated by Hansen (1999), is to use a consistent information criterion with a penalty for additional parameters big enough to eliminate overfitting. Admittedly, some researchers may prefer these kinds of alternatives to the post–sample forecast strategy. This study does not deny the usefulness of these other strategies; it simply focuses on the commonly used approach of comparing post–sample forecasts from the selected model to those from a benchmark model.

2. Data mining and forecast evaluation frameworks

2.1 Data mining framework

In practice, developing a model for a predictand y may involve several different forms of data mining. Some researchers start with a general model, including a variety of explanatory variables, and then sequentially drop variables to arrive at some best specification. The so-called LSE methodology described in, among others, Hendry (1995) and Mizon (1995) refines and extends this practice, subjecting each model specification to not only tests of the significance of individual variables but also tests of heteroskedasticity, stability, normality, etc.[7] Others, including Granger, King, and White (1995) and Hansen (1999), advocate using an information criterion such as the BIC to determine which of all candidate models is best.[8] Alternatively, researchers such as Ashley, Granger, and Schmalensee (1980) begin by estimating a general model and then drop, as a block, all variables with insignificant t-statistics. Still other researchers focus on finding a single variable x with explanatory power for y and search across a set of candidate x variables. Finally, apart from any data mining conducted in the course of a single study, the model specification adopted by a

[6] In this study's simulations, the Lovell (1983) rule of thumb proves effective for eliminating overfitting, consistently yielding empirical size of roughly α. But only an occasional study, such as Baba, Hendry, and Starr (1992), adjusts critical values as suggested by Lovell. In practice researchers typically just base results on 5 or 10 percent critical values, even though some sort of specification search has been conducted.

[7] Hendry and Krolzig (1999) and Krolzig and Hendry (2000) suggest refinements of LSE methodology, such as using more stringent critical values, designed to reduce the probability of overfitting.

[8] In addition, studies such as George and McCullough (1993) have developed Bayesian methods for model selection.

particular researcher may reflect the results of a profession–wide search.

The results presented below focus on a single form of data mining — a simple general–to–specific modeling procedure. This procedure begins with estimating a general model, regressing y on a total of K different x variables. The x variable with the smallest insignificant t–statistic is then dropped, and the model is reestimated. Variables are dropped one–by–one, based on the ordering of t-statistics, until all variables remaining in the model are significant. This general–to–specific procedure, which amounts to the stepwise regression described in Theil (1971), is a simplified version of the algorithm used in Hoover and Perez (1999). The more involved Hoover and Perez algorithm and the refinements developed in Hendry and Krolzig (1999) and Krolzig and Hendry (2000) are designed to address the ability of LSE methodology to arrive at a correct model. The simpler algorithm used here is designed to reflect the less sophisticated model search commonly conducted in practice.

While not presented in the interest in brevity, Monte Carlo results for several other model selection approaches are similar to the general–to–specific results reported below. Specifically, broadly similar results are obtained for algorithms that consist of: (1) searching across all possible combinations of variables to find the model that minimizes the BIC; (2) estimating a general model and then dropping, as a block, all variables with insignificant t-statistics; and (3) searching to find the single x with the greatest explanatory power for y.[9]

2.2 Forecast evaluation framework

Following McCracken (1999) and Clark and McCracken (2000), suppose a sample of observations $\{y_t, z_{1,t}^{'}\}_{t=1}^{T+1}$ that includes a scalar random variable y_t to be predicted and a $(k_0 + k_1 = k \times 1)$ vector of predictors $z_{1,t} = (z_{0,t}^{'}, z_{11,t}^{'})^{'}$. The sample is divided into in–sample and out–of–sample portions. Abstracting from the initial observations necessitated by the lags included in the esti-

[9] Admittedly, the probability of in–sample overfitting is lower for model selection based on the BIC than for the other approaches. But, consistent with the results in Hansen (1999), in small samples even the BIC has a considerable probability of overfitting.

mated models, the in–sample observations span 1 to R. Letting P denote the number of 1-step ahead predictions, the out–of–sample observations span $R+1$ through $R+P$.

Forecasts of y_t, $t = R+1, \ldots, R+P$, are generated using two linear models of the form $z_{i,t-1}^{'}\beta_i^*$, $i = 0, 1$, each of which is estimated. Model 1 is the specification selected by the general–to–specific search procedure. Model 0 is a restricted version of model 1; the exact form of this restricted model, which varies across experiments, is detailed in the next section. Under the null, model 1 nests the restricted model 0 and hence model 1 includes k_1 excess parameters. Under the alternative hypothesis, the k_1 restrictions are not true, and model 1 is correct.

The forecasts are *recursive*, 1-step ahead predictions. Under the recursive scheme, each model's parameters, β_i^*, $i = 0, 1$, are estimated with added data as forecasting moves forward through time: for $t = R+1, \ldots, R+P$, model i's prediction of y_t, $z_{i,t-1}^{'}\hat{\beta}_{i,t-1}$, is created using the parameter estimate $\hat{\beta}_{i,t-1}$ based on data from 1 to $t-1$. Generating forecasts by the *rolling* and *fixed* schemes considered in West and McCracken (1998) yields similar results. This analysis focuses on 1-step ahead forecasts because, as noted by McCracken (1999) and Clark and McCracken (2000), for multi-step forecasts, the asymptotic distributions of the forecast tests generally depend on the parameters of the data-generating process.

Forecasts from models 0 and 1 are compared using the following four tests: (1) the F–type test of equal MSE developed by McCracken (1999), denoted MSE-F; (2) the regression–based t–test for equal MSE proposed by Granger and Newbold (1977), labeled MSE-REG; (3) the ENC-NEW encompassing test developed by Clark and McCracken (2000); and (4) the regression–based t–test for encompassing proposed by Ericsson (1992), denoted ENC-REG.[10] The appendix provides detail on the computation of each statistic. Under the null that the restrictions imposed in model

[10] Clark and McCracken (2000) find the regression–based tests for equal MSE and encompassing perform slightly better than the analogous GMM–based t-tests proposed by Diebold and Mariano (1995) and Harvey, Leybourne, and Newbold (1998), respectively. In addition, simulations using Clark and McCracken's DGP-I indicate the finite–sample powers of the tests considered in this paper generally exceed the power of the out–of–sample test of causality developed by Chao, Corradi, and Swanson (2000).

0 are correct, the MSE for model 0 should be less than or equal to the MSE for model 1, and the encompassing regression coefficient should be less than or equal to 0. The MSE-F and MSE-REG statistics are compared against asymptotic critical values tabulated by McCracken (1999); the ENC-NEW and ENC-REG tests are compared against asymptotic critical values tabulated by Clark and McCracken (2000).

3. Simulation design

This study uses simulations to first evaluate the "size" performance of the model search procedure and out–of–sample forecast comparisons. In the size analysis, in truth the x regressands considered have no predictive power for the dependent variable. Each simulated data set is subjected to the general–to–specific search procedure, and in the event the data mining yields a model with at least one significant x regressand, out–of–sample forecasts from the selected model are compared to forecasts from a restricted model that corresponds to the true data-generating model. In this analysis, "size" measures the frequency with which the tests indicate some x variables have predictive power for the dependent variable, when there are no x variables in the DGP.

Simulations are then used to evaluate the "power" performance of the model search and out–of–sample forecast comparisons. In this analysis, in truth some x regressands have explanatory power for the dependent variable. Again, each simulated data set is subjected to a general–to–specific model search, and in the event the search yields a model including the x regressands that appear in the DGP, out–of–sample forecasts from the selected model are compared to forecasts from a restricted model. "Power" corresponds to the frequency with which the tests find that the x variables included in the DGP have predictive power.

This study relies on two different simulation approaches — simple Monte Carlo experiments and an alternative simulation framework patterned on those of Lovell (1983) and Hoover and Perez (1999). This section describes these approaches in turn.

3.1 Monte Carlo simulations

The DGP used in the Monte Carlo analysis takes the form

$$\begin{aligned} y_t &= .3y_{t-1} + b\ x_{1,t-1} + u_t \\ x_{i,t} &= .5x_{i,t-1} + v_{i,t}, \quad i = 1, \ldots, K, \end{aligned} \tag{1}$$

where K is the number of candidate explanatory variables (x) for y, and the error terms are all independently and identically distributed standard normal variables. In the size results, the coefficient b is set to 0. In the power results, in the interest of brevity b is set at just 0.2. Power results for other coefficient values and for DGPs incorporating two rather than just one x variable are qualitatively similar to those reported. In each Monte Carlo simulation, an initial observation is drawn from the unconditional normal distribution implied by the model parameterization and then $R+P$ observations are constructed using the autoregressive model structure and draws of the error terms from the standard normal distribution.[11]

In each simulation draw, the general–to–specific model selection procedure begins with an estimate of the regression

$$y_t = \alpha + \beta y_{t-1} + \sum_{i=1}^{K} \phi_i x_{i,t-1} + \epsilon_t. \tag{2}$$

This basic specification is designed with 1-step ahead prediction in mind.[12] As described in section 2, individual x variables are dropped sequentially until all remaining have significant t-statistics (a constant and one lag of y are always included in the estimated equation). The number of x variables considered, K, takes on a range of values: 3, 5, 10, and 20. In the reported results, models are selected using the significance levels most common in practice, 10% and 5%.

If one or more x variables remain in the model selected by the general–to–specific algorithm, out–of–sample forecasts of y are generated from the selected model (in section 2's notation, model

[11] Data are generated such that, for a given R, the in–sample data (the first R observations) are the same across experiments using different settings of P.

[12] Studies such as Stock and Watson (1998) and Knox, Stock, and Watson (2000) use the same basic regression.

1) using the last P observations of the sample. These forecasts are compared to predictions from an estimate of the restricted model (in section 2's notation, model 0)

$$y_t = \delta + \gamma y_{t-1} + e_t. \tag{3}$$

In the size experiments, the restrictions that model 0 imposes on model 1 are correct. In the power results, in which the coefficient b in (1) is non-zero, the restrictions are not true.

In the size results presented below, the model specification search is conducted using two different sample periods. The main set of results is generated by using just the first R observations in the model specification search, following the recommendation of Ashley, Granger, and Schmalensee (1980). Another set of results is generated by using all $R+P$ observations in the model specification search, following common practice. In this case, generating forecasts requires reestimating the forecasting models with just the first R observations and then iterating forward.

Results are reported for selected, empirically relevant combinations of P and R: $R = 100$ with $P = 20$, 40, and 100; and $R = 200$ with $P = 20$, 40, and 200.

3.2 LHP simulations

Additional evidence on whether post–sample forecast tests can help prevent data mining–induced overfitting is generated using an alternative simulation framework patterned on those of Lovell (1983) and Hoover and Perez (1999). In this framework, a general–to–specific model search is applied to an artificial dependent variable and a set of candidate regressands that are standard, quarterly macroeconomic data series. The artificial dependent variable y is generated using three different DGPs:

$$y_t = .267y_{t-1} + 3.460u_t \tag{4}$$

$$y_t = .267y_{t-1} + .263\Delta \ln PCE_{t-1} + 3.375u_t \tag{5}$$

$$y_t = .267y_{t-1} + .180\Delta \ln PCE_{t-1} + .031\Delta \ln DJIA_{t-1} + 3.306u_t \tag{6}$$

where u_t is a standard normal random variate, PCE denotes real personal consumption expenditures, and DJIA denotes the Dow Jones Industrial Average index deflated by the chain price index for GDP. The DGP (4) is drawn from an AR(1) estimated for quarterly GDP growth from 1959:3 to 2000:1. The DGPs (5) and (6) are drawn from regressions of GDP growth on lagged growth in GDP, consumption spending, and the DJIA, with the restriction that the coefficient on lagged GDP growth remain the same as in (4).

In each simulation draw, the general–to–specific model selection procedure begins with an estimate of the regression

$$y_t = \alpha + \beta y_{t-1} + \sum_{i=1}^{19} \phi_i x_{i,t-1} + \epsilon_t, \tag{7}$$

where, in this case, the x variables are the standard, quarterly macroeconomic data series listed in Table 1. These variables are the same as those considered by Hoover and Perez (1999) and essentially the same as those considered by Lovell (1983). As described above, the individual x variables are dropped sequentially until all remaining are statistically significant (a constant and one lag of y are always included in the estimated equation).

In this analysis, the data are again divided into in–sample and out–of–sample portions, using just the in–sample portion for model selection and reserving the out–of–sample portion for forecast evaluation. With the macroeconomic data series available from 1959:3 through 2000:1 (after differencing), R and P are set to 135 and 27, respectively. Accordingly, in the model selection procedure applied in each simulation, the sample period of the dependent variable spans 1959:4 through 1993:2. If one or more x variables remain in the model selected by the general–to–specific algorithm, forecasts are generated for 1993:3 through 2000:1. This sample split produces a ratio P/R for which McCracken (1999) and Clark and McCracken (2000) provide asymptotic critical values for the forecast test statistics considered.

In the "size" analysis, the DGP is model (4), and forecasts from the model determined by the

general–to–specific search (model 1) are compared to predictions from an estimate of the restricted specification (model 0)

$$y_t = \delta + \gamma y_{t-1} + e_t. \tag{8}$$

As described above, the comparison takes the form of testing whether forecasts from (8) have the same MSE as predictions from the general–to–specific–determined model and whether forecasts from (8) encompass predictions from the more general model. In the "power" analysis, the DGP is either model (5) or (6). In simulation draws in which the selected model includes some x variables, forecasts from the selected model are compared to projections from an estimate of the restricted model (8).

4. Results

Three key size results emerge from the Monte Carlo and LHP simulations.

4.1 Size

Size result 1. The Monte Carlo results in Table 2 confirm that data mining in the form of the model selection procedure used in this paper generally leads to overfitting. Consider, for example, the case in which the number of regressands included in the search (K) equals 10, and the 10% significance level is used in determining whether a variable belongs in the model. In these experiments, the search procedure yields an overfit model — a model including some x variables even though the DGP does not — in nearly 70% of the simulations (see the *in-sample* row of the third panel). As would be expected, for a given significance level, overfitting becomes more likely as the number of variables included in the search increases. When K rises to 20, empirical size rises to roughly 90% (see the *in-sample* row of the fourth panel). As would also be expected, for a given K, lowering the significance level used in model selection to 5% reduces the probability of overfitting.

The LHP simulation results in Table 3 provide further evidence that searches across real

macroeconomic data sets are likely to yield overfit models. In these simulations, when a 10% significance level is used in model selection, the general–to–specific search yields an overfit model in 87.3% of the simulations. Although using a more stringent 5% significance level improves matters somewhat, the probability of overfitting remains high, at 62.6%.

Size result 2. When just the in–sample portion of the data (the first R observations) are used in model selection, post–sample forecast evaluations can provide a very useful tool for avoiding overfitting. Conditioned on the selection procedure yielding a model that includes some regressands x with in–sample explanatory power, several of the post–sample forecast tests are close to being correctly sized.

More specifically, consider the Monte Carlo experiment in which the number of regressands included in the search (K) equals 10, the variable selection significance level is 10%, and $R = 100$ and $P = 20$. As reported in Table 2, in this experiment the MSE-REG and ENC-REG tests are roughly correctly sized: the tests indicate the selected regressands have predictive power in, respectively, 4.6% and 8.1% of the simulations in which the search procedure yields a model with some x regressands (see the first column of figures in the third panel). The MSE-F and ENC-NEW tests are subject to somewhat larger distortions, the latter more so than the former. In the same experiment, these statistics have size of 11.9% and 20.2%, respectively. The sizes of the tests remain largely unchanged as either K or the significance level used in the in–sample model selection change, but the size performance of the tests improves somewhat as the number of post–sample observations (P) increases.

LHP simulations confirm that post–sample forecast evaluations may be useful for avoiding overfitting in model searches involving typical macroeconomic data sets. For example, as shown in Table 3, both MSE-REG and ENC-REG are roughly correctly sized. When the model search uses the 10% level in determining whether a variable is significant, these two tests indicate the regressands in a selected model have predictive power in, respectively, 4.4% and 6.4% of the simulations

in which the search procedure yields a model with some x regressands.

Size result 3. As shown in Table 4, when the full sample of data is used in model selection, post–sample forecast evaluations are of limited help in avoiding overfitting. Using the full sample of data in the model search leads to a considerable chance that even forecast tests will spuriously indicate some x variables have predictive power. Consider again a Monte Carlo experiment in which the number of regressands included in the search (K) equals 10, the 10% significance level is used in variable selection, and $R = 100$ and $P = 20$. In this example, the MSE-F test indicates the selected regressands have predictive power in 40.6% of the simulations in which the full-sample search procedure yields a model with some x regressands (see the first column of figures in the third panel). The corresponding sizes of the MSE-REG, ENC-NEW, and ENC-REG tests are 22.4%, 50.6%, and 30.3%, respectively. The performance of the post–sample tests changes little as K — the dimension of the search — increases.

Out of concern that using the full sample of data in model selection might distort forecast–based inferences, Ashley, Granger and Schmalensee (1980) recommend using just in–sample data in the selection procedure and saving a portion of the data for out–of–sample forecast evaluation. Nonetheless, in practice, many researchers first use the full sample of data to select a model and then evaluate forecast performance. The results in Table 4 show that, under this approach, forecast comparison is unlikely to be an effective tool for avoiding overfitting.

4.2 Power

This section evaluates the frequency with which out–of–sample tests correctly determine that the x variables included in the data mining–determined model have predictive power for the dependent variable y. In particular, this section presents separate "power" results for the case in which the selected model corresponds to the DGP, which includes some x variables, and the case in which the selected model nests the DGP (the definition of nested here means that some of the models included in this category match the DGP exactly while others include some additional x

variables).[13] In both cases, "power" corresponds to the frequency with which the post–sample tests correctly find that the x variables included in the selected model have predictive power.

In order to provide a power benchmark, another out–of–sample test is added to the battery of tests considered. This additional statistic, denoted OOS GC, is simply a standard Granger causality test based on estimating the selected model using just the *out–of–sample* data. In particular, the OOS GC statistic is an F-test of exclusion restrictions on the x variables in the model yielded by the general–to–specific search, reestimated with just the out–of–sample data. In practice, of course, it is rare that a researcher conducts such a test (even though it is a very simple way of testing hypotheses without contamination from pre–test search), but in many respects this test provides a useful benchmark for the power of the forecast–based tests.[14]

Power result 1. The frequency with which the in–sample search procedure yields a model matching the DGP falls as the number of variables included in the search (K) increases. Consider, for example, Monte Carlo experiments with $R = 100$, $P = 20$, and a variable selection significance level of 10%. As shown in Table 5, the probability of the model search identifying the true model falls from 45.6% when $K = 5$ to 26.0% when $K = 10$ (see the *in-sample* entries of the first column, in the second and third panels). This finding conforms with what might be expected: the probability of identifying the true model falls as the dimension of the search expands.

Moreover, the frequency with which the in–sample search procedure yields a model matching the DGP typically rises as the significance level used in variable selection is reduced from 10% to 5%. Continuing with the same example, given $K = 10$, the probability of identifying the true model increases from 26.0% to 35.6% when the significance level used in model selection is reduced from 10% to 5%. But results not presented in the interest of brevity show that the effects of

[13] Results not reported in the interest of brevity confirm that, as expected, the probability of the in–sample model search yielding a completely "wrong" model — one that includes x variables but not the variable actually in the DGP — rises with K. Conditioned on model selection yielding a "wrong" specification, rejection rates for the post–sample forecast tests are broadly comparable to the sizes reported in Table 2.

[14] While not reported in the size results of Tables 2-4, the OOS GC test is about correctly sized.

using even more stringent significance levels are not monotone. Reducing the variable selection significance level to 1% raises the probability of selecting the true model in some cases and reduces the probability in other cases.

Power result 2. The probability of the in–sample search procedure yielding a model nesting the DGP remains unchanged as the number of variables included in the search rises, but falls as the variable selection significance level is reduced from 10% to 5%. Consider again the Monte Carlo example with $R = 100$, $P = 20$, and a variable selection significance level of 10%. As reported in Table 6, the probability of selecting a model nesting the DGP is roughly 71% for all values of K considered (see the *in-sample* entries of the first column, for panels 1-4). Because the probability of selecting a nesting model is unchanged but the probability of selecting the true model declines as K rises, the likelihood of overfitting increases with K. But reducing the significance level used in model selection from 10% to 5% causes the probability of selecting a nesting model to fall to about 60%. The more stringent significance level reduces overfitting.

Power result 3. The finite–sample powers of the post–sample forecast tests generally follow a simple ranking: ENC-NEW > MSE-F, ENC-REG > MSE-REG. These rankings apply to both of the in–sample model selection outcomes considered — the search yielding a model matching the DGP and nesting the DGP. For example, in Table 5's Monte Carlo results for $R = 100$, $P = 20$, $K = 10$, and a variable selection significance level of 10%, the ENC-NEW, MSE-F, ENC-REG, and MSE-REG test powers are 60.7%, 46.0%, 34.6%, and 20.8%, respectively (see the first column of figures in the third panel). So, in this example, the ENC-NEW and MSE-F tests correctly indicate the included x variables have predictive content in approximately one-half of those simulations in which the in–sample specification search yields a model matching the DGP. While the MSE-F and ENC-REG test powers do not follow a single ranking, the simulation results in Tables 5-7 show that, when P is small, the MSE-F test is typically more powerful. But when P is relatively large, the ENC-REG statistic is usually more powerful. For instance, when $R = 100$, $K = 10$, and the

variable selection significance level is 10%, but $P = 100$, the ENC-REG test has power of 79.1%, compared to 72.0% for MSE-F (see the third column in the third panel of Table 5).

Using the simple OOS GC test as a benchmark, the post–sample forecast tests appear to have good power. The powers of the ENC-NEW and ENC-REG tests consistently exceed that of the OOS GC statistic, and in most settings, the MSE-F test is also more powerful than the OOS GC test.[15] As an example, if model selection based on 5% significance levels yields the "true" model, the powers of the ENC-NEW, MSE-F, ENC-REG, and OOS GC tests are 63.6%, 46.4%, 34.7%, and 16.5%, respectively, when $R = 100$, $P = 20$, and $K = 5$ (see the seventh column in the second panel of Table 5). Similarly, if the selected model nests the DGP, the powers of the ENC-NEW, MSE-F, ENC-REG, and OOS GC tests are 60.8%, 43.1%, 32.7%, and 15.7%, respectively, when $R = 100$, $P = 20$, and $K = 5$ (see the seventh column in the second panel of Table 6).

Simulations based on the real–data LHP framework yield similar results. As shown in Table 7, for DGPs (5) and (6), the power of the ENC-NEW statistic is generally greater than the power of the MSE-F test, which in turn dominates the ENC-REG statistic and, last, the MSE-REG test. All of the forecast tests are more powerful than the OOS GC statistic. For example, using DGP (5) and a variable selection significance level of 5%, when the selected model matches the DGP the ENC-NEW, MSE-F, MSE-REG, ENC-REG, and OOS GC tests have power of 29.4%, 28.2%, 17.9%, 20.6%, and 8.0%.

While the reported power figures are not size–adjusted, the qualitative results — in particular, the general ranking of the tests and the finding of good power — are unlikely to be affected by size distortions. The MSE-F, MSE-REG, and ENC-REG tests are sufficiently close to being correctly sized that adjusting for size distortions would have only modest effects on power levels. Moreover, unreported Monte Carlo results for a simpler model search procedure that makes size adjustment

[15] In this paper's simulations, if the selected model matches the DGP, MSE-F is uniformly more powerful than OOS GC. If the selected model nests the DGP, MSE-F is more powerful when P is small but less powerful when P is large.

tractable yield the same basic power ranking of the tests.[16] This alternative search procedure consists of a simple, one–by–one search across the set of candidate x variables to find the single variable that has the greatest explanatory power for y (the biggest t–statistic). While a variety of factors would make size adjustment very difficult in the case of general–to–specific selection, size adjustment is relatively simple under the alternative form of data mining. Empirical critical values are generated in size experiments using the simpler model selection procedure, and then the empirical critical values from a given size experiment are used to compute adjusted power in the corresponding power experiment.

Power result 4. Finally, whether the dimension of the model search affects the powers of the post–sample tests depends on whether the selected model matches or nests the DGP. If the in–sample selection procedure yields the "true" model, the powers of the post–sample tests remain unchanged as the number of variables in the search (K) rises. For example, as shown in Table 5's Monte Carlo results for $R = 100$ and $P = 20$, when the selected model matches the DGP, the power of the MSE-F test is roughly 46% for all K values considered (see the first column of figures in panels 1-4). But if the selected model nests the DGP, the powers of the post–sample tests fall as K rises. As reported in Table 6's results for $R = 100$ and $P = 20$, when 10% significance levels are used in model selection, the power of the MSE-F test declines from 42.0% for $K = 3$ to 23.0% for $K = 20$ (see the first column of figures in the first and fourth panels). The decline in power in this case, as well as the difference in power between the matching and nesting cases, reflects a greater degree of overfitting.

4.3 Optimal sample split

For a researcher interested in using out–of–sample forecast comparisons to reduce the chances of data mining–induced overfitting, an important practical issue is the split of data into in–sample

[16] Without adjustment, power results for the alternative model selection procedure are very similar to those for the general–to–specific approach. As expected, adjusted power for the alternative selection procedure is somewhat lower than unadjusted power (more so for small P than large P). Moreover, size–adjusted power differences are somewhat smaller than the unadjusted power differences.

and out–of–sample portions. Unfortunately, however, the complexity of the issue rules out providing simple advice on how much data to reserve for post–sample forecast comparison. What is optimal will depend on how the researcher views the tradeoffs that exist and the sizes of the search and the sample.

The key tradeoff associated with the sample split is power. Monte Carlo results not reported in the interest of brevity indicate the sample split has little impact on size but significant effects on power. In these simulations, total sample sizes of $R + P = 160$ and $R + P = 400$ were each split six different ways, so as to produce P/R ratios between 0.2 and 2.0. In these experiments, the empirical "size" of the in–sample selection procedure is essentially unchanged as more of the data sample is used for forecasting, and less is used for in–sample model selection. But the "powers" of the in–sample selection procedure and the post–sample forecast tests do change with the sample split. As expected, as more of a data set is used for out–of–sample forecasts, the probability of the in–sample selection procedure correctly finding x has predictive power falls, while the powers of the post–sample forecast tests rise.

The auxiliary simulations conducted for this study show that using more data for forecasting sometimes, but not always, produces gains in forecast test power that exceed in–sample power loss.[17] The power effects of using more of a data sample for forecasting appear to depend on the particular forecast test, the dimension of the in–sample specification search, and the size of the sample. Given this finding and the reality that different researchers will have different objectives, it is impossible to provide any simple advice on the optimal sample split.

5. Application

This section uses an illustrative application to provide further evidence on the effectiveness of out–of–sample forecast comparisons in the presence of data mining. In particular, this section

[17] An additional difficulty in weighing power tradeoffs is that the test powers are not size–adjusted, as such adjustment would be very difficult.

uses the general–to–specific procedure described above to develop a model for core consumer price inflation in the United States, and then compares out–of–sample forecasts from the selected model to forecasts from a simple autoregressive model to determine if the selected explanatory variables in fact have predictive power for core inflation.

The inflation variable to be modeled is specified as a quarterly inflation rate less trend inflation. The trend, computed using Cogley's (1998) adaptive measure, captures low–frequency changes in monetary policy and inflation regimes. More specifically, inflation is measured using the chain price index for personal consumption expenditures (PCE) excluding food and energy. Letting π_t denote the log difference in the price index (in annualized percentage terms), detrended inflation is defined as $\pi_t - \bar{\pi}_{t-1}$, where, following Cogley, $\bar{\pi}_{t-1}$ is computed using exponential smoothing (with a smoothing coefficient of 0.125). Because the trend entering this inflation variable is lagged, 1-step ahead forecasts of detrended inflation are equivalent to 1-step ahead forecasts of actual inflation.

The model selection procedure begins with a general specification relating detrended inflation to a constant and four lags of: detrended inflation, relative food price inflation, relative energy price inflation, relative import price inflation, the log change in unit labor costs, and the rate of capacity utilization in manufacturing.[18] Food and energy prices are measured using chain price indexes for the food and energy components of PCE. Import prices are measured using the chain price index for total imports. Relative inflation rates for food, energy, and imports are defined as log changes in the corresponding price indexes less core PCE inflation. As described above, the model selection procedure begins with an estimate of the general model and then sequentially deletes variables with insignificant t–statistics, using a significance level of 5%.

The available data, which start in 1959:Q1, are divided into in-sample and out-of-sample portions so as to produce a modest P/R value for which McCracken (1999) and Clark and Mc-

[18] Watson (2000) points out that movements in capacity utilization parallel movements in the first factor index used for forecasting in Stock and Watson (1998, 1999) and Knox, Stock, and Watson (2000). In this application, replacing capacity utilization with unemployment produces very similar results.

Cracken (2000) report corresponding asymptotic critical values. After allowing for model lags and using roughly five years of data to initialize the trend series, the in–sample and out–of–sample periods are defined as 1965:Q1-1989:Q4 and 1990:Q1-1999:Q4, respectively. With this sample split, $R = 100$ and $P = 40$, so $P/R = 0.4$.

As shown in the upper panel of Table 8, in this application the in–sample selection procedure yields a model (denoted model 1) in which the explanatory variables are the first lag of the dependent variable, the second lag of relative energy price inflation, the first lag of relative import price inflation, and the fourth lag of capacity utilization. At least in–sample, this selected model fits the data better than the AR(1) model (denoted model 0) it nests.[19]

As reported in the lower panel of the table, the out–of–sample evidence indicates that relative energy and import price inflation and capacity utilization indeed have predictive power for inflation. The MSE-F, MSE-REG, ENC-NEW, and ENC-REG tests all reject the null of no predictive power. In this application, then, forecast comparisons show that the in–sample explanatory power of relative energy and import price inflation and capacity utilization is not a spurious result of the data mining used to select the model.

6. Conclusions

Building on McCracken (1999) and Clark and McCracken's (2000) recent work on the evaluation of forecasts from pre–specified nested models and Lovell's (1983) and Hoover and Perez's (1999) work on data mining, this paper examines the performance of out–of–sample forecast comparisons when the nesting model is the result of data mining. The basic results are drawn from simulations of a simple Monte Carlo design and a real data–based design similar to those in Lovell (1983) and Hoover and Perez (1999). The data mining takes the form of a general–to–specific modeling procedure. In each simulation, if the selected specification includes any of the candidate explanatory variables, tests of equal MSE and encompassing are applied to competing forecasts

[19] The use of one lag in model 0 minimizes both the AIC and SIC (for an AR model).

from the selected model and a benchmark model.

The simulations indicate out–of–sample forecast comparisons can help prevent overfitting. Most of the post–sample tests — which are only conducted if the model selection procedure indicates some of the considered variables have explanatory power — are roughly correctly sized. For forecast comparisons to be useful, though, they must be based on data not used in the specification search; as recommended by Ashley, Granger, and Schmalensee (1980), the data must be divided into in–sample and out–of–sample portions, with the latter reserved for forecast comparison. First using all of the data in the model search and then using an in–sample portion to reestimate the chosen model and an out–of–sample portion to generate forecasts is of relatively little help in avoiding overfitting. Finally, the post–sample forecast tests have relatively good power, although some are consistently more powerful than others.

The analysis concludes with an application, modeling quarterly U.S. inflation. In this example, a general–to–specific search procedure yields a model in which lagged energy price inflation, import price inflation, and capacity utilization have explanatory power for inflation. Tests of equal forecast accuracy and encompassing confirm the predictive content in energy and import price inflation and capacity utilization.

Appendix: The Forecast Test Statistics

Let $\hat{u}_{0,t} = y_t - z'_{0,t-1}\hat{\beta}_{0,t-1}$ and $\hat{u}_{1,t} = y_t - z'_{1,t-1}\hat{\beta}_{1,t-1}$ denote the 1-step ahead forecast errors for models 0 and 1, respectively, where model 1 is selected by the model search and model 0 is a restricted version of model 1. To simplify notation, for any variable w_t, let $\sum_t w_t$ denote $\sum_{t=R+1}^{R+P} w_t$.

The MSE-F and ENC-NEW tests are computed as follows:

$$\text{MSE–F} = P \cdot \frac{P^{-1}\sum_t \hat{u}_{0,t}^2 - P^{-1}\sum_t \hat{u}_{1,t}^2}{P^{-1}\sum_t \hat{u}_{1,t}^2} = P \cdot \frac{MSE_0 - MSE_1}{MSE_1} \tag{A1}$$

$$\text{ENC–NEW} = P \cdot \frac{P^{-1}\sum_t (\hat{u}_{0,t}^2 - \hat{u}_{0,t}\hat{u}_{1,t})}{P^{-1}\sum_t \hat{u}_{1,t}^2}. \tag{A2}$$

The MSE-REG test is simply the t-statistic associated with the coefficient α from the OLS regression (over the sample $R+1$ to $R+P$)

$$(\hat{u}_{0,t} - \hat{u}_{1,t}) = \alpha(\hat{u}_{0,t} + \hat{u}_{1,t}) + \text{error term}. \tag{A3}$$

Similarly, the ENC-REG test is the t-statistic for the coefficient α from the regression

$$\hat{u}_{0,t} = \alpha(\hat{u}_{0,t} - \hat{u}_{1,t}) + \text{error term}. \tag{A4}$$

As established by McCracken (1999) and Clark and McCracken (2000), the asymptotic distributions of the MSE-F, MSE-REG, ENC-NEW, and ENC-REG statistics do not depend on the parameters of the data-generating process, but do depend on two elements. The first is the number of restrictions model 0 imposes on model 1 (the number of excess variables in model 1 under the null, or k_1). The second is $\pi \equiv \lim_{P,R\to\infty} P/R$. For each combination of R and P considered, the simulations use asymptotic critical values from McCracken and Clark and McCracken for $\pi = \hat{\pi} \equiv P/R$.

REFERENCES

Amano, Robert A., and Simon van Norden, 1995, "Terms of Trade and Real Exchange Rates: The Canadian Evidence," *Journal of International Money and Finance* 14 (February), pp. 83-104.

Ashley, R., C.W.J. Granger, and R. Schmalensee, 1980, "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica* 48 (July), pp. 1149-67.

Baba, Yoshihisa, David F. Hendry, and Ross M. Starr, 1992, "The Demand for M1 in the U.S.A, 1960-1988," *Review of Economic Studies* 59 (January), pp. 25-61.

Campos, Julia, and Neil R. Ericsson, 1999, "Constructive Data Mining: Modeling Consumers' Expenditure in Venezuela," *Econometrics Journal* 2 (no. 2), pp. 226-40.

Chao, John, Valentina Corradi, and Norman Swanson, 2000, "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics*, forthcoming.

Chinn, Menzie D., and Richard A. Meese, 1995, "Banking On Currency Forecasts: How Predictable Is Change In Money?" *Journal of International Economics* 38 (February), pp. 161-178.

Clark, Todd E., and Michael W. Mccracken, 2000, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, forthcoming.

Cogley, Timothy, 1998, "A Simple Adaptive Measure of Core Inflation," Working Paper 98-6, Federal Reserve Bank of San Francisco, September.

Denton, Frank T., 1985, "Data Mining as an Industry," *Review of Economics and Statistics* 67 (February), pp. 124-27.

Diebold, Francis X., and Roberto S. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (July), pp. 253-63.

Ericsson, Neil R., 1992, "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration," *Journal of Policy Modeling* 14 (August), pp. 465-95.

Evans, Martin D.D., and Richard K. Lyons, 1999, "Order Flow and Exchange Rate Dynamics," *Journal of Political Economy*, forthcoming.

George, Edward I., and Robert E. McCulloch, 1993, "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association* 88 (September), pp. 881-89.

Granger, Clive W.J., Maxwell L. King, and Halbert White, 1995, "Comments on Testing Economic Theories and the Use of Model Selection Criteria," *Journal of Econometrics* 67 (May), pp. 173-87.

Granger, C.W.J., and Paul Newbold, 1977, *Forecasting Economic Time Series*, New York: Academic Press.

Hand, David J., 1999, "Discussion Contribution on 'Data Mining Reconsidered: Encompassing and the General–to–Specific Approach to Specification Search' by Hoover and Perez," *Econometrics Journal* 2 (no. 2), pp. 241-43.

Hansen, Bruce E., 1999, "Discussion of 'Data Mining Reconsidered'," *Econometrics Journal* 2 (no. 2), pp. 192-201.

Harvey, David I., Stephen J. Leybourne, and Paul Newbold, 1998, "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16 (April), pp. 254-59.

Hendry, David F., 1995, *Dynamic Econometrics*, New York and Oxford: Oxford University Press.

Hendry, David F., and Hans–Martin Krolzig, 1999, "Improving on 'Data Mining Reconsidered' by K.D. Hoover and S.J. Perez," *Econometrics Journal* 2 (no. 2), pp. 202-219.

Hoover, Kevin D., and Stephen J. Perez, 1999, "Data Mining Reconsidered: Encompassing and the General–to–Specific Approach to Specification Search," *Econometrics Journal* 2 (no. 2), pp. 167-191.

Knox, Thomas, James H. Stock, and Mark W. Watson, 2000, "Empirical Bayes Forecasts of One Time Series Using Many Predictors," manuscript, Harvard University.

Krolzig, Hans–Martin, and David F. Hendry, 2000, "Computer Automation of General–to–Specific Model Selection Procedures," *Journal of Economic Dynamics and Control*, forthcoming.

Lettau, Martin, and Sydney Ludvigson, 2000, "Consumption, Aggregate Wealth and Expected Stock Returns," *Journal of Finance*, forthcoming.

Lovell, Michael C., 1983, "Data Mining," *Review of Economics and Statistics* 65 (February), pp. 1-12.

Mccracken, Michael W., 1999, "Asymptotics for Out–of–Sample Tests of Causality," manuscript, Louisiana State University.

Mccracken, Michael W., 2000, "Data Mining and Out–of–Sample Inference," manuscript, Louisiana

State University.

Meese, Richard A., and Kenneth Rogoff, 1983, "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?" *Journal of International Economics* 14 (February), pp. 3-24.

Meese, Richard, and Kenneth Rogoff, 1988, "Was It Real? The Exchange Rate–Interest Differential Relation Over The Modern Floating–Rate Period," *Journal of Finance* 43 (September), pp. 933-948.

Mizon, Grayham E., 1995, "Progressive Modeling of Macroeconomic Time Series: the LSE Methodology," in *Macroeconometrics: Developments, Tensions and Prospects*, Kevin D. Hoover, ed., Boston: Kluwer, pp. 107-70.

Stock, James H., and Mark W. Watson, 1998, "Diffusion Indexes," NBER Working Paper #6702.

Stock, James H., and Mark W. Watson, 1999, "Forecasting Inflation," *Journal of Monetary Economics* 44 (October), pp. 293-335.

Theil, Henri, 1971, *Principles of Econometrics*, New York: Wiley.

Watson, Mark W., 2000, "Macroeconomic Forecasting Using Many Predictors," manuscript, Princeton University.

West, Kenneth D., 1996, "Asymptotic Inference About Predictive Ability," *Econometrica* 64 (September), pp. 1067-84.

West, Kenneth D., and Michael W. Mccracken, 1998, "Regression–Based Tests of Predictive Ability," *International Economic Review* 39 (November), pp. 817-40.

White, Halbert, 2000, "A Reality Check for Data Snooping," *Econometrica* 68 (September), 1097-1126.

Table 1
Macroeconomic Variables Used in Simulations

	Variable	Transformation
1.	GDP (1996$)	Δln
2.	Disposable personal income (1996$)	Δln
3.	Personal consumption expenditures (1996$)	Δln
4.	Gross private domestic investment (1996$)	Δln
5.	Government spending (1996$)	Δln
6.	Federal government spending (1996$)	Δln
7.	Federal government receipts, total	Δln
8.	Chain price index for GDP	$\Delta^2 ln$
9.	Composite index of coincident indicators	Δln
10.	Total reserves (adjusted for changes in reserve requirements)	Δln
11.	Monetary base (St. Louis Fed measure)	$\Delta^2 ln$
12.	M1	Δln
13.	M2	Δln
14.	Dow Jones Industrial Average (index)	Δln
15.	Moody's Aaa seasoned corporate bond yield	Δ
16.	Civilian labor force	Δln
17.	Civilian unemployment rate	Δ
18.	New orders (all manufacturing industries)	Δln
19.	Unfilled orders (all manufacturing industries)	Δln

Notes:
1. Variables 1-6 are chain-weight series.
2. Series 9-19 are constructed as within-quarter averages of the source monthly data.
3. Following Hoover and Perez (1999), the quantity variables available only in nominal terms — variables 7, 10, 11, 12, 13, 14, 18, and 19 — are deflated using the chain price index for GDP.
4. The integration orders reflected in the transformations reported in the last column generally match the orders used by Hoover and Perez (1999). The only differences are that, based on augmented Dickey-Fuller tests applied to the log series, variables 5, 7, 10, and 18 are treated as I(1) rather than I(2).
5. The raw data sample (prior to transformation) spans 1959:1 through 2000:1.

Table 2: Monte Carlo Results on Empirical Size
Model Selection Based on In–Sample Data
Nominal Size of Forecast Tests = 5%

	$R=100$			$R=200$			$R=100$			$R=200$		
	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$
	Nominal Size Used in Model Selection = 10%						**Nominal Size Used in Model Selection = 5%**					
$K=3$												
In-sample	.276	.276	.276	.277	.277	.277	.144	.144	.144	.151	.151	.151
MSE-F	.128	.098	.062	.148	.133	.058	.147	.096	.050	.160	.135	.056
MSE-REG	.054	.049	.027	.061	.047	.025	.051	.044	.021	.060	.042	.020
ENC-NEW	.199	.171	.154	.198	.195	.154	.244	.191	.186	.229	.208	.178
ENC-REG	.081	.081	.073	.081	.082	.066	.093	.086	.074	.082	.089	.070
$K=5$												
In-sample	.442	.442	.442	.418	.418	.418	.250	.250	.250	.235	.235	.235
MSE-F	.144	.105	.058	.151	.118	.046	.142	.109	.057	.153	.121	.046
MSE-REG	.062	.042	.026	.052	.045	.019	.058	.040	.021	.052	.044	.018
ENC-NEW	.229	.191	.154	.200	.178	.132	.246	.207	.170	.214	.191	.157
ENC-REG	.096	.081	.071	.074	.073	.058	.097	.082	.074	.075	.074	.060
$K=10$												
In-sample	.695	.695	.695	.673	.673	.673	.438	.438	.438	.416	.416	.416
MSE-F	.119	.080	.048	.133	.110	.052	.127	.086	.048	.150	.127	.061
MSE-REG	.046	.028	.020	.047	.041	.020	.047	.030	.019	.048	.043	.022
ENC-NEW	.202	.178	.154	.191	.183	.148	.227	.193	.167	.222	.208	.165
ENC-REG	.081	.071	.068	.065	.072	.067	.088	.080	.069	.073	.075	.074
$K=20$												
In-sample	.918	.918	.918	.891	.891	.891	.697	.697	.697	.661	.661	.661
MSE-F	.096	.063	.024	.124	.094	.032	.116	.071	.034	.145	.108	.038
MSE-REG	.032	.022	.010	.045	.031	.014	.037	.024	.014	.046	.035	.016
ENC-NEW	.202	.171	.151	.204	.194	.165	.224	.195	.176	.226	.212	.181
ENC-REG	.070	.063	.054	.072	.068	.064	.080	.065	.059	.077	.072	.067

Notes:
1. The DGP takes the form given in equation (1), with the coefficient b set to 0.
2. In each simulation, a general–to–specific procedure is used to identify the best model for y. The algorithm begins by estimating model (2) and then sequentially deletes any insignificant x variables until only significant variables remain.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. The significance level used in the model selection procedure is either 10% or 5%, and just the in–sample portion of the data is used in the model search.
5. In each simulation that the model yielded by the general-to-specific procedure includes at least one x variable, 1–step ahead out-of-sample forecasts are generated from the selected model and from an estimated AR(1) equation for y.
6. The *In-sample* row of the table reports the percent of the simulations in which the selected model includes at least one x variable. The remaining rows report the frequency with which, in simulations where the selected model includes at least one x variable, forecasts from the AR model are as accurate as or encompass those from the selected model. The post–sample test statistics are defined in section 2.2.
7. The number of simulations is 5000.

Table 3: LHP Simulation Results on Empirical Size Model Selection Based on In–Sample Data Nominal Size of Forecast Tests = 5%		
	Nominal Size Used in Model Selection	
	10%	5%
In-sample	.873	.626
MSE-F	.086	.100
MSE-REG	.044	.050
ENC-NEW	.140	.160
ENC-REG	.064	.072

Notes:
1. Each simulation uses artificial data for the dependent variable y and the 19 quarterly macroeconomic variables listed in Table 1 for the x regressands to be considered. The DGP for y is given in equation (4).
2. In each simulation, a general–to–specific procedure is used to identify the best model for y. The algorithm begins by estimating model (7) and then sequentially deletes any insignificant x variables until only significant variables remain.
3. With the 19 macroeconomic variables available from 1959:Q3 through 2000:Q1 (after transformation), the number of in–sample observations R and post–sample predictions P are set to 135 and 27, respectively, to obtain $\hat{\pi} = .2$.
4. The significance level used in the model selection procedure is either 10% or 5%, and just the in–sample portion of the data is used in the model search.
5. In each simulation that the model yielded by the general-to-specific procedure includes at least one x variable, 1–step ahead out-of-sample forecasts are generated from the selected model and from an estimated AR(1) equation for y.
6. The *In-sample* row of the table reports the percent of the simulations in which the selected model includes at least one x variable. The remaining rows report the frequency with which, in simulations where the selected model includes at least one x variable, forecasts from the AR model are as accurate as or encompass those from the selected model. The post–sample test statistics are defined in section 2.2.
7. The number of simulations is 5000.

Table 4: Monte Carlo Results on Empirical Size
Model Selection Based on Full Data Sample
Nominal Size of Forecast Tests = 5%

	$R=100$			$R=200$			$R=100$			$R=200$		
	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$
	Nominal Size Used in Model Selection = 10%						**Nominal Size Used in Model Selection = 5%**					
$K=3$												
In-sample	.278	.284	.277	.277	.276	.271	.148	.158	.144	.154	.145	.145
MSE-F	.349	.411	.463	.319	.360	.488	.405	.498	.572	.354	.437	.599
MSE-REG	.201	.265	.379	.176	.200	.375	.213	.307	.453	.181	.225	.441
ENC-NEW	.428	.494	.527	.380	.438	.545	.524	.620	.753	.434	.562	.778
ENC-REG	.254	.343	.454	.215	.256	.472	.294	.416	.582	.225	.307	.592
$K=5$												
In-sample	.437	.414	.424	.412	.429	.419	.237	.239	.232	.221	.239	.228
MSE-F	.358	.431	.472	.312	.359	.495	.416	.514	.577	.370	.430	.594
MSE-REG	.215	.278	.361	.158	.201	.386	.223	.322	.415	.179	.232	.444
ENC-NEW	.451	.525	.572	.377	.439	.561	.542	.662	.768	.454	.554	.776
ENC-REG	.277	.369	.467	.192	.253	.492	.311	.443	.581	.226	.306	.610
$K=10$												
In-sample	.686	.665	.672	.665	.657	.671	.430	.412	.419	.403	.404	.411
MSE-F	.406	.447	.519	.332	.370	.507	.453	.522	.591	.378	.446	.600
MSE-REG	.224	.277	.410	.169	.211	.385	.241	.313	.454	.188	.239	.440
ENC-NEW	.506	.563	.644	.414	.476	.630	.582	.674	.797	.479	.566	.804
ENC-REG	.303	.388	.538	.216	.284	.531	.347	.453	.620	.243	.333	.634
$K=20$												
In-sample	.898	.901	.895	.898	.889	.877	.679	.683	.674	.678	.662	.641
MSE-F	.427	.490	.592	.353	.418	.570	.470	.535	.643	.391	.459	.624
MSE-REG	.242	.313	.478	.167	.227	.455	.256	.331	.498	.183	.246	.475
ENC-NEW	.579	.696	.784	.464	.572	.762	.622	.737	.853	.496	.615	.843
ENC-REG	.357	.493	.673	.240	.336	.638	.371	.506	.698	.258	.345	.670

Notes:
1. The full sample of data ($R+P$ observations) are used in selecting the model.
2. See the notes to Table 2.

Table 5: Monte Carlo Results on Power — Selected Model is True Model
Model Selection Based on In-Sample Data
Nominal Size of Forecast Tests = 5%

	$R=100$			$R=200$			$R=100$			$R=200$		
	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$
	Nominal Size Used in Model Selection = 10%						**Nominal Size Used in Model Selection = 5%**					
$K=3$												
In-sample	.556	.556	.556	.761	.761	.761	.528	.528	.528	.809	.809	.809
MSE-F	.451	.532	.714	.509	.605	.905	.453	.534	.707	.516	.604	.903
MSE-REG	.217	.307	.530	.223	.319	.769	.207	.296	.508	.225	.315	.764
ENC-NEW	.601	.728	.902	.665	.795	.991	.624	.749	.915	.674	.802	.991
ENC-REG	.348	.500	.785	.346	.517	.961	.350	.511	.793	.353	.515	.960
OOS GC	.157	.281	.608	.160	.287	.888	.156	.288	.614	.163	.288	.888
$K=5$												
In-sample	.456	.456	.456	.616	.616	.616	.480	.480	.480	.723	.723	.723
MSE-F	.465	.525	.688	.504	.595	.905	.464	.518	.675	.507	.601	.904
MSE-REG	.216	.301	.505	.217	.306	.773	.209	.285	.478	.211	.309	.767
ENC-NEW	.615	.711	.898	.664	.775	.991	.636	.735	.909	.670	.787	.991
ENC-REG	.345	.490	.757	.340	.498	.960	.347	.494	.763	.338	.510	.961
OOS GC	.167	.278	.587	.157	.275	.884	.165	.276	.586	.158	.281	.889
$K=10$												
In-sample	.260	.260	.260	.341	.341	.341	.356	.356	.356	.537	.537	.537
MSE-F	.460	.516	.720	.506	.597	.904	.465	.524	.719	.505	.595	.902
MSE-REG	.208	.285	.512	.224	.322	.788	.217	.286	.513	.212	.314	.774
ENC-NEW	.607	.721	.897	.651	.787	.993	.634	.742	.914	.661	.785	.991
ENC-REG	.346	.479	.791	.336	.507	.962	.355	.495	.801	.333	.506	.962
OOS GC	.156	.278	.607	.150	.282	.885	.162	.280	.624	.149	.276	.886
$K=20$												
In-sample	.072	.072	.072	.108	.108	.108	.193	.193	.193	.301	.301	.301
MSE-F	.465	.546	.681	.510	.616	.909	.469	.550	.708	.510	.613	.904
MSE-REG	.238	.324	.493	.215	.306	.770	.243	.299	.513	.215	.309	.764
ENC-NEW	.612	.717	.911	.657	.788	.996	.620	.727	.914	.667	.786	.993
ENC-REG	.343	.518	.762	.347	.508	.970	.360	.524	.786	.346	.524	.963
OOS GC	.139	.299	.593	.152	.284	.898	.167	.298	.624	.163	.279	.888

Notes:
1. The DGP takes the form given in equation (1), with the coefficient b set to 0.2.
2. In each simulation, a general–to–specific procedure is used to identify the best model for y. The algorithm begins by estimating model (2) and then sequentially deletes any insignificant x variables until only significant variables remain.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. The significance level used in the model selection procedure is either 10% or 5%, and just the in–sample portion of the data is used in the model search.
5. In each simulation that the model yielded by the general-to-specific procedure matches the DGP for y, 1–step ahead out-of-sample forecasts are generated from this selected model and from an estimated AR(1) equation for y.
6. The *In-sample* row of the table reports the percent of the simulations in which the selected model matches the DGP for y. The remaining rows report the frequency with which, in those simulations where the selected model corresponds to the DGP, forecasts from the AR model are as accurate as or encompass those from the selected model. The forecast test statistics are defined in section 2.2. The OOS GC test is an F-test of Granger causality based on just the post–sample data.
7. The number of simulations is 5000.

Table 6: Monte Carlo Results on Power — Selected Model Nests True Model
Model Selection Based on In-Sample Data
Nominal Size of Forecast Tests = 5%

	$R=100$			$R=200$			$R=100$			$R=200$		
	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$	$P=20$	$P=40$	$P=100$	$P=20$	$P=40$	$P=200$
	Nominal Size Used in Model Selection = 10%						**Nominal Size Used in Model Selection = 5%**					
$K=3$												
In-sample	.710	.710	.710	.942	.942	.942	.598	.598	.598	.898	.898	.898
MSE-F	.420	.493	.662	.489	.577	.882	.435	.508	.674	.504	.588	.890
MSE-REG	.197	.276	.482	.212	.298	.741	.196	.276	.482	.218	.301	.746
ENC-NEW	.586	.705	.886	.647	.775	.988	.614	.732	.905	.664	.790	.990
ENC-REG	.324	.470	.756	.332	.495	.952	.337	.490	.773	.344	.503	.956
OOS GC	.151	.266	.580	.155	.276	.879	.153	.277	.599	.160	.284	.886
$K=5$												
In-sample	.717	.717	.717	.938	.938	.938	.607	.607	.607	.890	.890	.890
MSE-F	.405	.442	.594	.459	.536	.850	.431	.471	.620	.485	.565	.868
MSE-REG	.185	.240	.419	.185	.262	.709	.190	.248	.425	.196	.280	.725
ENC-NEW	.572	.667	.860	.622	.740	.983	.608	.704	.886	.649	.766	.985
ENC-REG	.311	.432	.704	.302	.456	.941	.327	.462	.730	.322	.483	.948
OOS GC	.153	.246	.542	.143	.251	.854	.157	.260	.560	.151	.269	.871
$K=10$												
In-sample	.709	.709	.709	.938	.938	.938	.599	.599	.599	.890	.890	.890
MSE-F	.347	.373	.522	.408	.473	.774	.382	.426	.587	.446	.519	.818
MSE-REG	.149	.188	.349	.162	.230	.633	.175	.221	.404	.172	.254	.679
ENC-NEW	.529	.630	.822	.576	.709	.967	.579	.681	.862	.620	.745	.982
ENC-REG	.276	.382	.669	.276	.411	.909	.306	.425	.721	.296	.452	.933
OOS GC	.133	.227	.514	.132	.231	.821	.145	.244	.562	.138	.250	.852
$K=20$												
In-sample	.712	.712	.712	.923	.923	.923	.605	.605	.605	.878	.878	.878
MSE-F	.230	.242	.320	.329	.369	.633	.315	.342	.446	.401	.462	.738
MSE-REG	.096	.122	.214	.120	.163	.497	.134	.166	.303	.149	.210	.575
ENC-NEW	.438	.546	.740	.523	.635	.943	.516	.623	.813	.594	.702	.969
ENC-REG	.210	.302	.556	.232	.343	.856	.262	.382	.645	.266	.408	.895
OOS GC	.122	.205	.436	.123	.201	.773	.137	.234	.517	.135	.232	.827

Notes:
1. The DGP takes the form given in equation (1), with the coefficient b set to 0.2.
2. In each simulation, a general–to–specific procedure is used to identify the best model for y. The algorithm begins by estimating model (2) and then sequentially deletes any insignificant x variables until only significant variables remain.
3. R and P refer to the number of in–sample observations and post–sample predictions, respectively.
4. The significance level used in the model selection procedure is either 10% or 5%, and just the in–sample portion of the data is used in the model search.
5. In each simulation that the model yielded by the general-to-specific procedure nests the DGP for y, 1–step ahead out-of-sample forecasts are generated from this selected model and from an estimated AR(1) equation for y.
6. The *In-sample* row of the table reports the percent of the simulations in which the selected model nests the DGP for y. The remaining rows report the frequency with which, in those simulations where the selected model nests the DGP, forecasts from the AR model are as accurate as or encompass those from the selected model. The forecast test statistics are defined in section 2.2. The OOS GC test is an F-test of Granger causality based on just the post–sample data.
7. The number of simulations is 5000.

Table 7: LHP Simulation Results on Power Model Selection Based on In-Sample Data Nominal Size of Forecast Tests = 5%				
	Nominal Size Used in Model Selection			
	10%	5%	10%	5%
Selected Model is True Model				
	DGP(5)		*DGP(6)*	
In-sample	.081	.194	.035	.060
MSE-F	.281	.282	.514	.510
MSE-REG	.185	.179	.237	.265
ENC-NEW	.274	.294	.699	.715
ENC-REG	.215	.206	.399	.430
OOS GC	.101	.080	.121	.114
Selected Model Nests True Model				
	DGP(5)		*DGP(6)*	
In-sample	.565	.509	.280	.195
MSE-F	.165	.210	.332	.395
MSE-REG	.083	.112	.172	.192
ENC-NEW	.249	.278	.563	.625
ENC-REG	.121	.148	.306	.350
OOS GC	.067	.072	.090	.106

Notes:
1. Each simulation uses artificial data for the dependent variable y and the 19 quarterly macroeconomic variables listed in Table 1 for the x regressands to be considered. The DGPs used for y are given in equations (5) and (6).
2. In each simulation, a general–to–specific procedure is used to identify the best model for y. The algorithm begins by estimating the model (7) and then sequentially deletes any insignificant x variables until only significant variables remain.
3. With the 19 macroeconomic variables available from 1959:Q3 through 2000:Q1 (after transformation), the number of in–sample observations R and post–sample predictions P are set to 135 and 27, respectively, to obtain $\hat{\pi} = .2$.
4. The significance level used in the model selection procedure is either 10% or 5%, and just the in–sample portion of the data is used in the model search.
5. In each simulation that the model yielded by the general-to-specific procedure either matches or nests the DGP for y, 1–step ahead out-of-sample forecasts are generated from the selected model and from an estimated AR(1) equation for y.
6. In the upper panel, the *In-sample* row reports the percent of the simulations in which the selected model matches the DGP for y. The remaining rows report the frequency with which, in those simulations where the selected model corresponds to the DGP, forecasts from the AR model are as accurate as or encompass those from the selected model. The forecast test statistics are defined in section 2.2. The OOS GC test is an F-test of Granger causality based on just the post–sample data.
7. In the lower panel, the *In-sample* row reports the percent of the simulations in which the selected model nests the DGP for y. The remaining rows report the frequency with which, in those simulations where the selected model nests the DGP, forecasts from the AR model are as accurate as or encompass those from the selected model.
8. The number of simulations is 5000.

Table 8
Modeling Core Inflation: Estimates and Forecast Comparisons

In–Sample Model Estimates, 1965:Q1 to 1989:Q4 ($R = 100$)		
Explanatory variable	*Dependent variable* $= \pi_t - \bar{\pi}_{t-1}$	
	Model 0	**Model 1**
$\pi_{t-1} - \bar{\pi}_{t-2}$	.786 (.062)	.395 (.072)
$Relative\ energy\ price\ inflation_{t-2}$		.016 (.008)
$Relative\ import\ price\ inflation_{t-1}$		.042 (.009)
$Capacity\ utilization_{t-4}$		.087 (.018)
SEE	.905	.717
$\bar{R}^2$	.615	.758
Out–of–Sample Tests of Predictive Power, 1990:Q1 to 1999:Q4 ($P = 40$)		
	Test statistics	*5% Asymptotic critical values*
MSE (RMSE), Model 1	.438 (.662)	
MSE (RMSE), Model 2	.360 (.600)	
MSE-F	8.635	2.062
MSE-REG	1.092	.968
ENC-NEW	12.092	1.865
ENC-REG	3.057	1.529

Notes:
1. The dependent variable (and predictand) $\pi_t - \bar{\pi}_{t-1}$ is inflation in the core chain price index for consumption (π_t) in quarter t less lagged trend inflation. Section 5 of the text explains the variables in detail.
2. As detailed in Section 5, Model 1 is the result of applying a general–to–specific modeling procedure, beginning with a general model relating $\pi_t - \bar{\pi}_{t-1}$ to lags of itself and lags of food price inflation, energy price inflation, import price inflation, growth in unit labor costs, and capacity utilization in manufacturing.
4. The significance level used in the model selection procedure is 5%, and just the in–sample portion of the data is used in the model search.
5. The figures in parentheses in the upper panel of the table are OLS standard errors for the reported coefficient estimates.
6. The forecast results in the lower panel are based on 1–step ahead out-of-sample predictions of $\pi_t - \bar{\pi}_{t-1}$ from models 0 and 1. The forecast test statistics are defined in section 2.2. The asymptotic critical values for the MSE-F and MSE-REG tests are taken from McCracken (1999); the asymptotic critical values for the ENC-NEW and ENC-REG tests are from Clark and McCracken (2000).