Journal of Economic Methodology 7:2, 217–230 2000

I am not, nor have I ever been a member of a data-mining discipline

Clinton A. Greene

Abstract This paper argues classical statistics and standard econometrics are based on a desire to meet scientific standards for accumulating reliable knowledge. Science requires two inputs, mining of existing data for inspiration and new or 'out-of-sample' data for predictive testing. Avoidance of data-mining is neither possible nor desirable. In economics out-of-sample data is relatively scarce, so the production process should intensively exploit the existing data. But the two inputs should be thought of as complements rather than substitutes. And we neglect the importance of out-of-sample testing in the production of reliable knowledge. Avoidance of data-mining is not a substitute for tests conducted in new samples. The problem is not that data-mining corrupts the process, the problem is our collective neglect of out-of-sample encompassing, stability and forecast tests. So the data-mining issue diverts us from the crucial margin.

Keywords: repeated testing, stability, time series, experimental, data-mining, applied methods, prediction

1 INTRODUCTION

If methodology were an experimental science then one would be content to allow those who subscribe to different views on data-mining to proceed apace, the proof of the superior approach to be revealed in the productivity of each approach over time. If it is possible to use logical argument to clarify or settle the issue it is because one agrees on standards. In particular standards must be agreed upon for constructing a reliable basis for knowledge. This discussion presumes our ideal is to meet the standards of experimental science. This is the only reasonable context for adopting the assumptions of standard econometrics and I believe is the source of our aspirations, frustrations, and doubts in empirical investigation.

Hoover and Perez (2000) propose three attitudes towards data-mining. Truncating the attributes of these attitudes a bit, data-mining is something to be avoided. Or it is unavoidable. Or it is desirable. All three positions are taken by reputable economists who subscribe to a scientific ideal. And all statements are reasonable. Data-mining is to be avoided, is unavoidable and is desirable. This apparently self-contradictory statement is reasonable because

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online
http://www.tandf.co.uk/journals

the term 'data-mining' is vague with respect to content and with respect to context. Because it is vague and destructive rather than constructive I think Hoover and Perez are correct to attack the term. In fact I will argue that those economists who believe they can or should agree with the title of this essay are either pure speculative theorists, deluded or dishonest.

This paper takes the position that reliable knowledge requires experiment or its equivalent. Therefore a useful benchmark is that of classical statistics in which single tests are conducted in distinct samples. This is the lesson taught (usually) indirectly in econometrics courses in many forms, including the warning that evaluating the significance of a coefficient assumes all other coefficients are correct (joint confidence contours are not defined by individual confidence intervals) and via noting that pre-testing invalidates tests, so only a single test is statistically meaningful. But testing is not the only useful benchmark of the scientific approach to knowledge because experiments are explorations and suggest new 'specifications'. The process of scientific learning is data-mining in the broadest sense. The problem in applied time series work is to determine how learning and testing can validly or usefully co-exist.

The following discussion will begin by clarifying the contexts in which data-mining is desirable, why it is unavoidable and the sense in which it is possible and desirable to avoid data-mining. But the vagueness of the term invites misuse to such a degree one hopes in this paper to encourage the use of other terminology. It is more useful to think on the one hand about learning from the data and design criteria for models, and on the other hand to think about predictive testing. In fact one will argue that the last two words of the previous sentence are redundant. It is not possible to test in either the statistical or scientific sense except via work with new or out-of-sample data.

I am making three claims for most of applied time series econometrics. First, when interpreting test statistics calculated for time series which extend into the past, there is no distinction between a model estimated without specification search and a model designed via an explicit specification search. The 'data-mining' engaged in by one researcher is only a marginal contribution to the collective search process. Second, specification search and the use of in-sample test statistics are useful techniques which allow us to learn from the past and to avoid the repetition of mistakes in model specification. Third, the only cure for our discomforts and the only way to apply a scientific mentality in economics is to take data-inspired models and test them out-of-sample. This leads one to conclude the data-mining issue diverts us from the real weakness in applied economics and from the real weakness in the development of meaningful theory. The crucial weak-link is our apparent lack of commitment to (redundantly) scientific, predictive, precisely interpretable, statistical tests.

2 DATA-MINING IS DESIRABLE

Data-mining is a virtuous activity in the chain of scientific method: Collect information (experimental or not), contemplate, grapple with and mine the information to inspire a new idea, then test in new data (predict). The outcome of the test and data-mining of the new sample is then used to suggest a new experiment and test. The danger in the term 'data-mining' is that it attacks learning or taking inspiration from the data (the world) as well as the element of sloppy statistical testing. Positive investigation must wilt if the charge is taken seriously in all its possible breadth. The accusation in economic discourse is as dangerous as red-baiting is in politics.[1]

Hoover and Perez (2000) have produced a simulation study that examines the general-to-specific within-sample data-mining model specification methodology for auto-correlated (but not integrated) variables. The approach is scientific in the purest sense. The data is generated experimentally and controlled to vary in an empirically relevant manner. They find the approach works well in the sense of picking up the causal correlations and dropping variables that are not causal. They determine some of the factors which influence how well the method works. If their results can be replicated by others a few times then the usefulness of general to specific (G-S) specification will be settled for the case of non-integrated variables which can be divided into causal and non-causal variables.

It should not be surprising that specification search, which lets the data speak for itself, is an empirically useful strategy. If one is interested in empirical evidence then in time-series work most of the evidence consists of the old and much examined series which stretch into the past. However imperfect, one must use the evidence at hand. I argue below that such evidence has nothing to do with testing in any statistical or scientific sense, even if the researcher does not engage in specification search and conducts only one 'test'. But at a minimum a model is on tenuous grounds if it does not fit the past data or does not take account of important aspects of that data. That it may be easy to construct models which appear to have desirable characteristics in the current and past data is not a good justification for constructing models which cannot meet this limited standard.

3 DATA-MINING IS TO BE AVOIDED (WHEN CONDUCTING TESTS)

In the experimental context testing in conjunction with specification search is to be avoided and can be avoided. The data, stylized facts or regression results can and should be explored for puzzles, suggestive correlations or interesting outcomes to suggest a hypothesis. This mining of the sample data does not corrupt the test because a new set (or ideally, many new sets) of data will be experimentally generated for testing the hypothesis. The new data set is not

mined and must not be mined for the test to be convincing. The hypothesis simply stands or falls. Exploring the data for inspiration is a separate operation from using data for the test, the firewall being distinct data sets. Both links in the chain are essential to the scientific method.

To illustrate our difficulties in economics let us invent an unfortunate mythological creature, the classical time-series econometrician who is also a scientist, pure of heart. Our poor unfortunate happens to have data for only one planet and a few decades. As our classical econometrician would have it (and the classics are always right), a theoretician wakes up one morning to recall a dream which is the inspiration for a new theory.[2] The time series econometrician then translates the theory into a single regression model and translates a test of the theory into the outcome of a single statistic-based test.

But here the process must stop. If the outcome of the test is used to inspire revision of the theory then the theory cannot be tested. Because the existing data series is no longer free of pre-testing or specification search and so cannot yield test statistics with known distributions. An attempt to re-use the original data implies the actual distribution of any test statistic differs from standard distributions in an unknown manner. Since the sample data suggested the new theory or specification one already knows the sample lends some sort of support to the new theory. Unless one is sure that the time-series among different countries are truly independent (and there are reasons to believe they are not) then testing using another country's data does not escape this problem. Even if the data is independent between countries and the data is of useful quality for 200 of the world's nations this allows for only 200 legitimate tests (a small number per researcher) and the possibilities are quickly exhausted.

From this perspective data-mining refers to invalid statistical testing as a result of naive over-use of a sample. In particular, the use of a sample both for learning-inspiration and for testing of that which was learned or mined from the sample. Any test of a theory or model is corrupted if the test is conducted using data which overlaps that of any previous empirical study used to suggest that theory or model.

The moral is clear. Scientific and reliable knowledge requires experiments. And good econometrics free of distorted statistic distributions requires repeated investigation of data which is at least new, even if not experimentally controlled. Given the difficulty in economics of testing in an unexamined data set it would then seem that the corollary to 'avoid data-mining' is 'stick to pure theory', a position taken by many.[3]

But testing in un-mined data sets is a difficult standard to meet *only* to the extent one is impatient. There is a simple and honest way to avoid invalid testing. To be specific, suppose in 1980 one surveys the literature on money demand and decides the models could be improved. File the proposed improvement away until 2010 and test the new model over data with a starting date of 1981. In this simple approach the development and testing of theory would be constrained to move at a pace set by the rate at which new data

becomes available (at which experiments are conducted). That the development of science must be constrained by the pace of new experimentation seems obvious enough. The mistake is to suppose every new regression is a new experiment. Only new data represents a new experiment.

I do not consider this a pessimistic outlook. This is because I thinks much can be learned from exploring a sample. Patience and slow methodical progress are virtuous. And the impatient can conduct forecast based stability and encompassing tests with only a few years of new data. But seeing economists behave as though they do not believe in the central role of constraints and of inputs in the production of reliable knowledge, certainly is grounds for pessimism.

4 DATA-MINING IS UNAVOIDABLE

The 'data-mining' engaged in by one researcher is only a marginal contribution to the collective search process. This search process corrupts and invalidates all formal statistical 'tests' conducted with data extending backwards into the past. Anyone who takes the results of an empirical paper as a basis for further empirical research has engaged in collective data-mining.[4] If the data-sets overlap then data-mining is a problem in the sense of pre-testing and violation of assumptions necessary for statistical testing.

As an example, suppose one reads Goldfeld's (1973) investigation of money regressions. His results apply to the sample 1952–1973. If one then specifies a money regression which includes that period then all of his results on significance, fit and stability are pre-testing relative to your own study. This may be conscious as when ones uses his results to determine your choice of interest rate or unconscious in the sense that his study becomes part of the collective wisdom of informed researchers.

If one is to use overlapping samples then the only way to eliminate data-mining in the sense of pre-testing is to ban all reading of the empirical journals. To maintain the purity of investigations then for each distinct dependent variable one must allow a researcher one empirical paper containing only one regression and one test statistic.[5] For an empirical model to be entirely free of data-mining it must be based on a theory which is constructed without prior knowledge of the world. Maintaining the purity of the stock of potential doctoral candidates requires that all empirical examples be removed from graduate and undergraduate classes.

Even maintaining a stock of cloistered and pure (read uneducated) economists to avoid data-mining by individual researchers cannot eliminate collective distortion of statistical 'tests'. Suppose one-hundred cloistered economists wake up from a dream with new theories of inflation each of which involves distinct variables. They run a regression incorporating their new dream-inspired variable.[6] Under the null, each can consider their test statistic distributions to be undistorted and conventional. Some find results which are

positive, others negative. Those with positive results submit to journals and a subset of these is published. The result is spurious literature.[7]

That the accusation 'data-mining' deserves ridicule (as in the above paragraphs) is meant in all seriousness. Claiming that data-mining is always bad when in fact learning from the data is by necessity a part of any aspiration to scientific investigation is a recipe for fatalism and cynicism. At best the term is too vague to be useful and at worst promotes an anti-scientific spirit.

Simply stated, the problem is to determine how we can learn from our own and others' explorations of empirical data. Any learning is itself a specification search which explores and settles on some subset of the possibilities, so the problem is how one can one usefully conduct, report and evaluate specification searches. I believe the only reasonable approach to evaluation lies in efforts to separate specification search and data exploration from testing. All other routes diverge from basic scientific practice and are futile efforts to get something for free. But the limited size and availability of 'new' data sets in economics makes it all the more important to extract as much information as possible from the existing data before engaging in out-of-sample testing.

For time-series econometrics the production of knowledge requires two activities, exploring of one sample and testing in another. The supply of out-of-sample data is relatively scarce and is mostly driven by the passage of time. So the production process must exploit the out-of-sample data only when more cannot be learned from exploring the previous data. The out-of-sample data must be held in reserve for tests for which there is no substitute in the original data. In the production of knowledge good data-mining exploits the relatively abundant resource and conserves the scarce resource.

5 ENCOMPASSING AS A DESIGN CRITERIA FOR MODELS

If one accepts the argument that in-sample test statistics can never be used for formal tests of time-series models then how should one interpret the pervasive use of such statistics? This paper holds that the useful answer hinges on the notion of encompassing. The idea of encompassing is to be distinguished from formal encompassing statistics because encompassing criteria can be quite broad. But it will be useful to consider the G-S approach to learning from the data as the application of a narrowly defined encompassing criteria.

There is no reason to do empirical work unless the sample can be taken as representative of the population. The general-to-specific approach to designing models begins with a large set of variables (including lags) and begins to simplify the model. The criteria for simplification is quite simple. At each stage the variable with the lowest (in absolute value) t-statistic is dropped. This is equivalent to dropping the variable with the lowest F-statistic (for the restriction $b=0$). At each round the degrees of freedom for the F-statistic of each variable are the same, so this is equivalent to dropping the variable which

causes the least increase of the sum of squared residuals. At each round the question is 'If we were to prefer a marginally simplified model, which variable is least important in explaining (fitting) this sample?'. Among the models that include one less variable, the G-S method chooses the model which fits better than (encompasses) the other models.

Even if the *t*- or *F*-statistics were distributed as in the standard tables this would be irrelevant to the choice of how far to go in simplifying the model. The convention of choosing a significance level at which the restriction is rejected does not avoid applying one's tastes (1%, 5% or 10%) in test size. And this convention avoids the real question, which is the economic significance of the deterioration in the fit of the model at each round of simplification. Since conventional practice has yet to seriously recognize the importance of economic significance or to create economically meaningful standards for parsimony it is impossible to know precisely where the process of model simplification should stop.

The truth of the restrictions imposed in model simplification is not an issue and it is a mistake to think of *t*- or *F*-tests as tests of the truth. One always knows the correct answer to the question 'is a zero coefficient a reasonable hypothesis?'. The answer is no. The measure of the real line is not affected by the removal of the point '0' and the probability of any point hypothesis being correct is zero. From an economic perspective it is also a rare case in which b=0 is possible. In an Arrow-Debreu economy demand in one market is always a function of prices of every commodity indexed by every possible date. Imperfect competition increases the data needed to specify completely demand in a market by a huge factor over that needed in an Arrow-Debreu economy. To change the context, if one posits the hypothesis that anticipated money does not affect real output then in principal the problem is that every agent will have different access to and costs of acquiring information. Thus any particular approximation or measure of 'anticipated money' will be somewhat contaminated by including 'unanticipated money' for some subset of actors. So although one can hope the coefficient on 'anticipated money' is small, one already knows it is not zero.

At the same time one either prefers simple empirical models (as a matter of taste) or is forced to construct them because the data sets are a few observations shy of infinite. Thus in principle the list of correct or 'true' variables for a regression model is too large. In specification search one is not interested in the truth of '$b = 0$' or '$b \neq 0$', the truth here is known ($b \neq 0$, but it may be that $b = 0.00001$). One is interested first in ranking variables by their relative importance and second in whether a small-sample point estimate is so unreliable as to make a forecast with $b = 0$ (or perhaps $b = 1$, or $b = 0.5$) imposed a more reliable approach than forecasting with the small-sample point estimate.

The G-S approach lets the world (the sample) speak for itself in two respects. First, at each stage of simplification the evidence is used to determine

which variables are most important to retain. One might prefer a more or less parsimonious model, but one can not argue (on empirical grounds) against the evidence that G-S follows in determining the relative importance of the variables. Second, the criteria for parsimony in G-S specification is quite transparent. Always simplify the model until at some conventional significance level the simplification (restriction) is rejected. Since I regard all empirical models as exhaustively data-mined via collective effort I sees this as no less meaningful than applying any other test in-sample. The use of a *t*- or *F*-test is simply a conventional and arbitrary way to define how much parsimony is too much parsimony.

Thus I see G-S as setting up a hierarchy of nested models each of which is more parsimonious than the last and encompasses all others within its class. If one begins with twenty variables (including lags) then at each stage G-S chooses the best nineteen-variable model, the best eighteen-variable model and so on. To argue against this ordering is to argue that the sample cannot be trusted, in which case any empirical investigation is fruitless. And one can have confidence in the ordering while being tentative about whether the five or the twelve variable model provides the preferable degree of parsimony. Good data-mining uses the sample evidence to determine which variables are most important, and this is what G-S does.

I believe similar reasoning can be applied to most specification diagnostics. In the case of statistics which evaluate residual autocorrelation this is direct since the measure of such correlation will increase as lagged variables are dropped in model simplification. For statistics which measure the closeness of the distribution of residuals to normality a decision must be made about 'how normal' the residuals should look. Among models which contain 10 variables one can ask which among them appears to have the 'most normal' residuals, likewise for nine variables etc. Positing heterogeneous errors with a specific structure amounts to a claim that forecasts which take account of this structure, will encompass (be more accurate than) forecasts from a simple least-squares model. Within a sample one can compare the gains from various schemes of weighted-least-squares and apply one's tastes for parsimony. But in all cases formal tests can be conducted only with new samples.

Thus although it is somewhat reassuring that Hoover and Perez find in their study of G-S that nominal probability values for test statistics are usually close to actual values, I do not think this is central to the problem of specification search. Specification search uses the existing empirical evidence to rank models, especially those of equal complexity. This ordering is more important than cardinal measure. Thus I think that when reading Hoover and Perez's study of the G-S approach the key question is whether variables less important (economically significant) in population are usually dropped before variables which are more important in population.

Decisions about the proper degree of parsimony should depend on the use to which the model will be put. Specification search can employ statistics to

rank the costs of various coefficient restrictions without relying on formal tests of the statistical significance of such costs. In sample 'tests' should be interpreted as providing sample evidence of such rankings rather than as true statistical tests.

The alternative to collective and individual data-mining is to ignore the lessons of the past in designing models. It is of course always most convincing if one can show these lessons are usually a good guide to the future as in predictive testing, hence the emphasis on re-estimation and out-of-sample tests. But a pure theory-inspired model which happens to perform well in a new data set but does less well in the past data is defective. Such a case implies the model is specialized to the new data and thus of questionable reliability. Expecting theories and models to explain the current and past data imposes a minimal standard for empirical generality and disallows the generation of a new theory or model to explain each new event.

6 THE FEAR OF DATA-MINING

This paper has argued that the problem of distorted test statistic distributions is not curable for any model estimated with data extending into the past, the problem can be avoided only in predictive testing with new data. But there are other possible objections to data-mining. A legitimate fear discussed in Mayer (2000) is that results may be unconsciously or consciously manipulated, as when one variable is excluded because the exclusion is needed to ensure the value or nominal statistical significance of a key variable. A cure for outright dishonesty is provided by studies that seek to replicate results. But this would be redundant if models and results were more often re-applied to new or out-of-sample data. Since I see this as the key to settling any question in a reliable manner, the issue of manipulation becomes minor. Honest and self-conscious work is always to be preferred. But honest work need not provide reliable results. The only check is via out-of-sample confirmation.

And dishonesty (conscious and unconscious) is not cured by avoiding data-mining approaches. The distinction between the G-S methodology and manipulation of results is that in pure G-S the data speaks for itself. The more collective wisdom, prior beliefs and/or the judgement of the researcher are used to modify and shape specification search then the more important out-of-sample testing becomes. But this is because such shaping (such as imposing a priori coefficient restrictions) may violate the evidence embodied in the past data. The dishonesty inherent in manipulation of auxiliary variables to ensure a result for a variable of interest is likely to be revealed if one takes a pure data-mining approach as in G-S and compares the result to the manipulated specification.

Another reason to be uncomfortable with data-mining is that it can be seen as contributing to a proliferation of competing empirical models. The concern with proliferation of models is legitimate, but to see honest, non-manipulated

data-mining as contributing the to problem is a misplaced apprehension. Suppose three of four researchers posit the correct regressor in a model to be different and each has theoretical justification. And each estimates a model including only their preferred variable. But a fourth researcher begins by including all three variables. And suppose this fourth agnostic finds one of the variables statistically insignificant and drops it from the model. If one reads the papers of the first three researchers one will not know which model is best or whether all are complements. And so we might reasonably wish to see some sort of encompassing evidence or the results of estimating a nesting model. The fourth researcher has already done this.

It is the lack of attention to 'theory encompassing' and the lack of efforts to encompass prior empirical results that allows a proliferation of models. The field becomes balkanized with a variety of approaches too numerous to compare and too numerous for one researcher to solidly retain in memory as the previous state of knowledge. Under these circumstances simple lessons learned in the past are forgotten and knowledge is not cumulative over time. The solution to this problem is more attention to encompassing tests both within and especially out of sample.

Since G-S data-mining can be interpreted as using the encompassing principle to sort out the strengths and weaknesses of nested models it is part of the solution rather than part of the problem. And the problem of data-mining must be considered within the larger context of encompassing and not just data-based encompassing. Theory-mining is as much an enemy of progressive knowledge as is data-mining. If theories are not usually expected to prove themselves by being generalizations of older theories then models can proliferate endlessly, constrained only by the power of imagination. Intensive mining of previous knowledge both theoretical and empirical with the aim of meeting the encompassing standard is necessary if empirical work is to progressively increase economic knowledge.

7 SOME POSITIVE SUGGESTIONS

My greatest discomfort with data-mining, such as G-S, comes out of the fact it is mostly the dynamics which are data-mined and the dynamics generate most of the fit and forecasting ability of such models. In most error-correction models the error-correction or long-run term can be dropped with little loss in the accuracy of one-year-ahead forecasts. Yet we do not take the dynamics seriously in the sense of regarding the dynamics revealed by data-mining as a 'stylized fact' which stands as a theoretical puzzle or challenge. The stance of many is that 'theory has little to say about dynamics'. But if this is true then theory has little to say about or learn from the world.

A serious obstacle to theoretical interpretation of G-S dynamic models (and likewise ADL or VAR models) lies in allowing subsets of the variables that determine the long-run relationship to have independent effects. More

precisely, suppose the long-run relationship in an error-correction model is given by

$$y = LR = \alpha_1 x_1 + \alpha_2 x_2$$

Ignoring lags of the dependent variable, the G-S approach to specification of the error-correction model will begin with the ADL model (with lags 0 through l)

$$\Delta y_t = \Sigma b_{x1,l} \Delta x_{1,t-l} + \Sigma b_{x2,l} \Delta x_{2,t-l} - b_0(y_{t-1} - LR_{t-1}). \quad (1)$$

The lags of the differenced variables retained in the final G-S specification may have no relation to the long-run equilibrium. The variables $\Delta x_{1,t-l}$ may be dropped entirely, or the lags retained may not correspond to those of the other determinant of the long-run (x_2). A specification which gives more weight to theory is

$$\Delta y_t = \Sigma \beta_l \Delta LR_{t-l} - b_0(y_{t-1} - LR_{t-1}). \quad (2)$$

Equation 2 is much easier to interpret from a theoretical point of view. If the model is stable out-of-sample then one can make the case the model reveals something fundamental about adjustment costs as a function of rates of change and acceleration as in Salmon (1982). And the revealed dynamics are then a puzzle, which deserves theoretical interpretation and justification. Equation 1 is more general in that it allows different speeds of adjustment depending on the source of the change in the long-run equilibrium. But this is not a good argument for creating a model from which is difficult to learn any lessons. The point is to use empirical models which can create puzzles which stand as credible and compelling challenges for theoretical interpretation and development. Only if (1) proves to be superior to (2) is extra complication warranted. For the comparison to be made the simpler model must be considered, particularly in out-of-sample stability and encompassing tests.

As a consumer of the LSE approach I think some confusion is created by using the term 'test' both for in-sample specification criteria based on test statistics and actual statistical testing out-of-sample. It falsely suggests to the reader that practitioners believe they are testing when practitioners are in fact well aware they are simply applying design criteria. Since the reader knows data-mining distorts tests the term should probably be avoided and some other terminology invented when discussing specification search. Given the current state of applied work and the difficulty of changing norms in practice and in publishing it makes sense to address the most fundamental problems. The aversion to specification search will diminish and the benefits of any specification approach will be more compellingly revealed if out-of-sample testing and encompassing standards (formal and informal, empirical and

theoretical) are applied more routinely in empirical work. After all, general-to-specific simplifications simply apply the encompassing principal to a baseline model. Thus if one wants to promote specification search I think focusing mostly on these other two aspects of the LSE approach is a logical pre-condition and is strategically useful.

Currently there is too little commitment to either the encompassing standard or to serious (out-of-sample) testing. The most important proscription of science and statistics and the LSE school is to test out of the original sample (predict).[8] Yet updates of previous studies are rare in our journals. If we were committed to valid statistics and progressive knowledge then the acceptance of any applied paper would include an implicit commitment to publish future updates. To name some names, a rare counter-example (almost) is the update of Baba *et al.* (1992) produced by Hess *et al.* (1994; 1998). That this rare opportunity was not immediately picked up by a journal is evidence against the profession's commitment to scientific standards and understanding of basic statistics.[9] I think the best way to dislodge the fiction that tests conducted in old data are valid if the model is free of explicit specification search and the way to dislodge the notion that data-mined specifications are unreliable is to test such competing models in periodic updates using new data.

In fact this is the case to such a degree that in the interest of promoting data-mining in the long run it could be strategic to accept in the short run more resistance among editors and referees to data-mined specifications. But only *if* this was accompanied by an increase in the publication of updates, re-tests, and out-of-sample encompassing tests. The relative strength of data-mined specifications and the weakness of models which ignore lessons from the past data will be most convincing if revealed in out-of-sample testing. This is the only way to shift the debate from an a priori arena into an empirical and statistically precise arena.

8 CONCLUSION

This paper claims the issue of data-mining is a red-herring in three respects. First, all empirical work is and should be the result of some sort of data-mining. Given the scarcity of new data one must exploit the existing data as thoroughly as possible. In work which uses time-series data there is always a large component of collective data-mining embodied in the adopted form of the regression model. Statistics from data-mined specifications provide informal but valuable evidence or suggestions. The claim that one specification is less data-mined than another is not sufficient to justify formal interpretation of regression statistics as in classical statistics. All are guilty and a measure of explicit data-mining does not discriminate between useful and un-useful work. In-sample 'tests' are useful as design criteria but only out-of-sample tests are precisely meaningful applications of statistics. Without out-

of-sample testing there is no distinction between running regressions and constructing historical (ex post) narratives.[10] Second, if the applied journals were full of papers applying encompassing criteria and if for every previously published model there were one or two subsequent studies applying out-of-sample tests to that model then concern with data-mining would largely vanish. Likewise, the advantages of any particular approach to specification would be revealed. The passage of a few years time is sufficient to conduct forecast stability and forecast encompassing tests out-of-sample. If one believes in some variety of specification search as a productive research strategy then in the long run it is most essential to promote the use of encompassing and out-of-sample tests in the evaluation of models.

Finally, discomfort with data-mining is a misplaced reaction to other problems. Under current practice it is often difficult to interpret empirical work as contributing to progressive knowledge. One fear is that more pervasive or explicit data-mining will swamp us with mutations too numerous to classify, much less interpret and evaluate. The cure for this is the promotion of encompassing standards in the broadest sense. But the most important fear of data-mining stems from legitimate doubts about the validity of most testing in over-worked time series data and from the false hope that if our own contribution to data-mining is small then our own research will somehow circumvent the problem of pre-test distortion of standard statistic distributions. But avoidance of explicit specification search will not cure our discomfort with sloppy in-sample 'tests' nor will it insulate us from dishonest or deluded results. Specification search is the only way to learn from the data. Out-of-sample testing of data-mined specifications is the only way to conduct statistically valid tests and create reliable knowledge. All else represents a wish for scientific validity unconstrained by scientific inputs, technology, method and sensibility.

Clinton A. Greene
University of Missouri
clinton-greene@umsl.edu

NOTES

1 Hence the title of this essay.
2 Or the inspiration could come from interactions with other theorists, but only (as discussed below) if they in turn were isolated from all empirical information.
3 Applied economists are self-selected and perhaps their willingness to embrace impurity is grounds enough for ignoring them. On the other hand logic unrestrained by empirical confirmation is able to produce an endless supply of theories. Thus pure 'theory-mining' is as much a dead end as pure 'data-mining'.
4 Thus one is both a miner and a red!
5 In addition to producing a pure and reliable applied literature this would cure the problem of an excess demand for referees, excess demand for journal space and a distortionary inflation of requirements for quantity in publishing.
6 They omit money as a variable because they are well aware their choice of

monetary aggregate would be influenced by previous examinations of the same inflation data employed in their regressions.

7 If one advocates the view that new models should make sense from the perspective of currently available data then you are advocating data-mining.

8 Replication is not the issue because in economics by 'replication' we mean checking of calculations using the original researcher's data. But replication in the scientific sense means repeated generation of new samples under reproduced conditions, i.e. re-testing in new data.

9 Hess *et al*. (1998) unfortunately did not expect a receptive audience at the journal of original publication and so did not submit there on the first round. But if we were committed to meaningful testing then the editors of a number of journals would have solicited the submission of this all too rare presentation of out-of-sample tests.

10 An honorable and difficult task but unscientific nonetheless.

REFERENCES

Baba, Y., Hendry, D. F. and Starr, R. M. (1992) 'The demand for M1 in the USA, 1960–1984', *Review of Economic Studies* 59: 25–61.

Goldfeld, S. (1973) 'The demand for money revisited', *Brookings Papers on Economic Activity* 3: 577–646.

Hess, G. D., Jones, C. S. and Porter, R. D. (1998) 'The predictive failure of the Baba, Hendry and Starr Model of the demand for M1 in the United States', *Journal of Economics and Business* 50: 477–507.

Hoover, K. D. and Perez, S. J. (2000) 'Three attitudes towards data mining', *Journal of Economic Methodology* 7: 195–210.

Mayer, T. (2000) 'Data-mining: a reconsideration', *Journal of Economic Methodology* 7: 183–94.

Salmon, M. (1982) 'Error correction mechanisms', *Economic Journal* 92: 615–29.