Journal of Economic Methodology 7:2, 195–210 2000

Three attitudes towards data mining

Kevin D. Hoover and Stephen J. Perez

Abstract 'Data mining' refers to a broad class of activities that have in common, a search over different ways to process or package data statistically or econometrically with the purpose of making the final presentation meet certain design criteria. We characterize three attitudes toward data mining: first, that it is to be avoided and, if it is engaged in, that statistical inferences must be adjusted to account for it; second, that it is inevitable and that the only results of any interest are those that transcend the variety of alternative data mined specifications (a view associated with Leamer's extreme-bounds analysis); and third, that it is essential and that the only hope we have of using econometrics to uncover true economic relationships is to be found in the intelligent mining of data. The first approach confuses considerations of *sampling distribution* and considerations of *epistemic warrant* and, reaches an unnecessarily hostile attitude toward data mining. The second approach relies on a notion of robustness that has little relationship to truth: there is no good reason to expect a true specification to be robust alternative specifications. Robustness is not, in general, a carrier of epistemic warrant. The third approach is operationalized in the general-to-specific search methodology of the LSE school of econometrics. Its success demonstrates that intelligent data mining is an important element in empirical investigation in economics.

Keywords: data mining, extreme-bounds analysis, specification search, general-to-specific, LSE econometrics

1 INTRODUCTION

To practice data mining is to sin against the norms of econometrics, of that there can be little doubt. That few have attempted to justify professional abhorrence to data mining signifies nothing, few have felt any pressing need to justify our abhorrence of theft either. What is for practical purposes beyond doubt needs no special justification; and we learn that data mining is bad econometric practice, just as we learn that theft is bad social practice, at our mothers' knees as it were. Econometric norms, like social norms, are internalized in an environment in which explicit prohibitions, implicit example and, many subtle pressures to conformity mold our *morés*. Models of 'good' econometric practice, stray remarks in textbooks or lectures, stern warnings from supervisors and referees, all teach us that data-mining is

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online © 2000 Taylor & Francis Ltd
http://www/tandf.co.uk/journals

abhorrent. All agree that theft is wrong, yet people steal and, they mine data. So, from time to time moralists, political philosophers and legal scholars find it necessary to raise the prohibition against theft out of its position as a background presupposition of social life, to scrutinize its ethical basis, to discriminate among its varieties, to categorize various practices as falling inside or outside the strictures that proscribe it. Similarly, the practice of data mining has itself been scrutinized only infrequently (e.g., Leamer 1978; 1983, Mayer 1980, 1993; Lovell 1983; Hoover 1995). In this paper, we wish to characterize the practice of data mining and three attitudes towards it. The first attitude is the one that that we believe is the most common in the profession namely, data mining is to be avoided and, if it is engaged in, we must adjust our statistical inferences to account for it. The second attitude is that data mining is inevitable and that the only results of any interest are those that transcend the variety of alternative data mined specifications. The third attitude is that data mining is essential and that the only hope that we have of using econometrics to uncover true economic relationships is to be found in the intelligent mining of data.

2 WHAT IS DATA MINING?

'Data mining' refers to a broad class of activities that have in common a search over different ways to process or package data statistically or econometrically with the purpose of making the final presentation meet certain design criteria. An econometrician might try different combinations of regressors, different sample periods, different functional forms, or different estimation methods in order to find a regression that suited a theoretical preconception, had 'significant' coefficients, maximized goodness-of-fit, or some other criterion or set of criteria. To clarify the issues, consider a particularly common sort of data mining exercise. The object of the search is the process that generates $\mathbf{y}$, where $\mathbf{y} = [y_t]$, an $N \times 1$ vector of observations, $t = 1, 2, \ldots N$. Let $\mathbf{X} = \{X_j\}$, $j = 1, 2, \ldots M$, be the universe of variables over which a search might be conducted. Let $\mathbf{X}^P = \mathrm{X}\mathbf{X}$, the power set of $\mathbf{X}$ (i.e., the set of all subsets of $\mathbf{X}$). If $\mathbf{y}$ were generated from a linear process, then the actual set of variables that generated it is an element of $\mathbf{X}^P$. Call this set of true determinants $\mathbf{X}_T \in \mathbf{X}^P$, and let the true data-generating process be:

$$\mathbf{y}^k = \mathbf{X}_T^k \boldsymbol{\beta}_T + \boldsymbol{\omega}^k, \tag{1}$$

where $\omega^k = [\omega_t^k]$, the vector of error terms, and k indicates the different realizations of both errors and the variables in $\mathbf{X}$. Now, let $\mathbf{X}_i \in \mathbf{X}^P$ be any set of variables; these define a model:

$$\mathbf{y}^k = \mathbf{X}_i^k \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i^k, \tag{2}$$

where $\boldsymbol{\varepsilon}^k = [\varepsilon_t^k]$ includes ω^k, as well as every factor by which equation (2) deviates from the true underlying process in equation (1). Typically, in

economics only one realization of these variables is observed, k is degenerate and takes only a single value. In other fields, for example in randomized experiments in agriculture and elsewhere, k truly ranges over multiple realizations, each realization it is assumed, coming from the same underlying distribution. While in general, regressors might be random (the possibility indicated by the superscript k on $\mathbf{X}_i$), many analytical conclusions require the assumption that $\mathbf{X}_i$ remains fixed in repeated samples of the error term.[1] This amounts to, $\mathbf{X}_i^k = \mathbf{X}_i^h$, $\forall k, h$, while, $\varepsilon^k \neq \varepsilon^h$, $\forall k \neq h$, except on a set of measure zero.

We can estimate the model in equation (2) for a given i and any particular realization of the errors (a given k). From such estimations we can obtain various sample statistics. For concreteness, consider the estimated standard errors that correspond to $\hat{\boldsymbol{\beta}}_i$ the estimated coefficients of equation (2) for specification i.[2] What we would like to have are the population standard errors of the elements of $\hat{\boldsymbol{\beta}}_i$ about $\bar{\hat{\boldsymbol{\beta}}}_i$. Conceptually, they are the dispersion of the sampling distribution of the estimated coefficients while $\mathbf{X}_i$ remains fixed in repeated samples of the error term. Ideally, sample distributions would be calculated over a range of k's, and as k approached infinity, the sample distributions would converge to the population distributions. In practice there is a single realization of ε^k. While conceptually this requires a further assumption that the errors at different times are drawn from the same distribution (the ergodic property), the correct counterfactual question remains: what would the distribution be if it were possible to obtain multiple realizations with fixed regressors? Conceptually, the distribution of sample statistics is derived from repeatedly resampling the residual within a constant specification. This is clear in the case of standard errors estimated in Monte Carlo settings or from bootstrap procedures.[3] In each case, simulations are programmed that exactly mimic the analysis just laid out.

Data-mining in this context amounts to searching over the various $\mathbf{X}_i \in \mathbf{X}^P$ in order to meet selection criteria: e.g., that all of the t-statistics on the elements of $\mathbf{X}_i$ be statistically significant or that R^2 be maximized.

3 ONLY OUR PREJUDICES SURVIVE

Data mining is considered reprehensible largely because the world is full of accidental correlations, so that what a search turns up is thought to be more a reflection of what we want to find than what is true about the world. A methodology that emphasizes choice among a wide array of variables based on their correlations is bound to select variables that just happen to be related in the particular data set to the dependent variables, even though there is no economic basis for the relationship. One response to this problem is to ban search altogether. Econometrics is regarded as hypothesis testing. Only a well specified model should be estimated and if it fails to support the hypothesis, it fails; and the economist should not search for a better specification.

A common variant of this view, however, recognizes that search is likely and might even be useful or fruitful. However, it questions the meaning of the test statistics associated with the final reported model. The implicit argument runs something like this: Conventional test statistics are based on independent draws. The tests in a sequence of tests on the same data used to guide the specification search are not necessarily independent. The test statistics for any specification that has survived such a process are necessarily going to be 'significant'. They are 'Darwinian' in the sense that only the fittest survive. Since we know in advance that they pass the tests, the critical values for the tests could not possibly be correct. The critical values for such Darwinian test statistics must in fact be much higher. The hard part is to quantify the appropriate adjustment to the test statistics.

The interesting thing about this attitude towards data mining is the role that it assigns to the statistics. In the presentation of the textbook interpretation in the last section, those statistics were clearly reflections of *sampling distribution*. Here the statistics are proposed as *measures of epistemic warrant*, that is, as measures of our justification for believing a particular specification to be the truth or as measures of the nearness of a particular specification to the truth.[4] Sampling distribution is independent of the investigator: it is a relationship between the particular specification and the random errors thrown up by the world; the provenance of the specification does not matter. Epistemic warrant is not independent of the investigator. To take an extreme example, if we know an economist to be a prejudiced advocate of a particular result and he presents us with a specification that confirms his prejudice, our best guess is that the specification reflects the decision rule - search until you find a confirming specification.

Not all search represents pure prejudice but, if test statistics are conceived of as measures of epistemic warrant, the standard statistics will not be appropriate in the presence of search. Michael Lovell (1983) provides an example of this epistemic approach to test statistics. He argues that critical values must be adjusted to reflect the degree of search. Lovell conducts a number of simulations to make his point. In the first set of simulations, Lovell (1983: 2–4) considers a regression like equation (2) in which the elements of $\mathbf{X}$ are mutually orthogonal. He considers sets of regressors that include exactly two members (i.e., a fairly narrow subset of $\mathbf{X}^P$). The dependent variable $\mathbf{y}$ is actually purely random and unrelated to any of the variables in $\mathbf{X}$ (i.e., $\mathbf{X}_T = \varnothing$). Using five per cent critical values, he demonstrates that one or more significant *t*-statistics occur more than five percent of the time. He proposes a formula to correct the critical values to account for the amount of search.

Lovell (1983: 4–11) also considers simulations in which there are genuine underlying relationships and the data are not mutually orthogonal. He uses a data set of twenty actual macroeconomic series as the universe of search $\mathbf{X}$. Subsets of $\mathbf{X}$ (i.e., $\mathbf{X}_T$ for the particular simulation) with at most two members

are used to generate a simulated dependent variable. He then evaluates the success of different search algorithms in recovering the particular variables used to generate the dependent variable. These algorithms are different methods for choosing a 'best' set of regressors as $\mathbf{X}_i$ ranges over the elements of $\mathbf{X}^P$? Success can be judged by the ability of an algorithm to recover $\mathbf{X}_T$. Lovell also tracks the coefficients on individual variables $X_g \in \mathbf{X}$, noting whether or not they are statistically significant at conventional levels. He is, therefore, able to report empirical type I and type II error rates (i.e., size and power). As in his first simulation, he finds that there are substantial size distortions, so that conventional critical values would be grossly misleading. What is more, he finds low empirical power, which is related to the algorithms inability to recover $\mathbf{X}_T$. The critical point for our purposes is that Lovell's simulations implicitly interpret test statistics as measures of epistemic warrant. The standard critical value or the size of the test refers to the probability of a particular *t*-statistic on repeated draws of ω^k (*k* taking on multiple values) from the *same* distribution (that is the significance of the textbook assumption that the regressors are fixed in repeated samples). Lovell's experiment, in contrast, takes the error term in the true data-generating process, ω^k, to be fixed (there is a single *k* for each simulation) but, considers the way in which the distribution of ε_i^k, the estimated residual for each specification considered in the search process, varies with every new $\mathbf{X}_i$. Lovell's numbers are correct but the question they answer refers to a particular *application* of a particular search *procedure* rather than to any property of the *specification* independently in relation to the world.

The difficulty with interpreting test statistics in this manner is that the actual numbers are specific for a particular search procedure in a particular context. This is obvious if we think about how Lovell or anyone would conduct a Monte Carlo simulation to establish the modified critical values or sizes of tests. A particular choice must be made for which variables appear in $\mathbf{X}$ and a particular choice procedure must be adopted for searching over elements of $\mathbf{X}^P$. Furthermore, one must establish a measure of the amount of search and keep track of it. Yet, typically economists do not know how much search produced any particular specification, nor is the universe of potential regressors well defined. We do not start with a blank slate. Suppose, for example, we estimate a 'Goldfeld' specification for money demand (Goldfeld 1973; also Judd and Scadding 1982). How many times has it been estimated before? What do we know in advance of estimating it about how it is likely to perform? What is the range of alternative specifications that have been or might be considered? A specification such as the Goldfeld money demand equation has involved literally incalculable amounts of search. Where would we begin to assign epistemically relevant numbers to such a specification?

4 ONLY THE ROBUST SHOULD SURVIVE

Edward Leamer (1978, 1983; and in Hendry *et al*. 1990) embraces the implication of this last question. He suggests immersing empirical investigation in the vulgarities of data mining in order to exploit the ability of a researcher to produce differing estimates of coefficient values through repeated search. Only if it is not possible for a researcher to eliminate an empirical finding should it be believed. Leamer is a Bayesian. Yet, Bayesian econometrics present a number of technical hurdles that prevent even many of those who, like Leamer, believe that it is the correct way to proceed in principle from applying it in practice. Instead, Leamer suggests a practicable alternative to Bayesian statistics: extreme bounds analysis. The Bayesian question is, how much incremental information is there in a set of data with which we might update our beliefs? Leamer (1983) and Leamer and Leonard (1983) argue that, if econometric conclusions are sensitive to alternative specifications, then they do not carry much information useful for updating our beliefs. Data may be divided into *free variables*, which theory suggests should be in a regression; *focus variables*, a subset of the free variables which are of immediate interest; and *doubtful variables*, which competing theories suggest might be important.[5] Leamer suggests estimating specifications that correspond to every linear combination of doubtful variables in combination with all of the free variables (including the focus variables). The *extreme bounds* of the effects of the focus variables are given by the endpoints of the range of values (± 2 standard deviations) assigned to the coefficients on each of them across these alternative regressions. If the extreme bounds are close together then there can be some consensus on the import of the data for the problem at hand; and if the extreme bounds are wide, that import is not pinned down very precisely. If the extreme bounds bracket zero, then the direction of the effect is not even clear. Such a variable can be regarded as not *robust to alternative specification*.[6]

The linkage between extreme bounds analysis and Bayesian principles is not, however, one-to-one in the sense that the central idea, robustness to alternative specification, represents an attitude to data-mining held by non-Bayesians as well. Thomas Mayer's (1993, 2000) argument that every regression run by an investigator, not just the final preferred specification, ought to be reported arises from a similar notion of robustness. If a coefficient is little changed under a variety of specifications, we should have confidence in it, and not otherwise. Mayer's proposal that the evidence ought not to be suppressed, but reported, at least in a summary fashion (e.g., as extreme bounds) is, he argues, an issue of honest communication and not a deep epistemological problem. But we believe that this is incorrect. The epistemological issue is this: if all the regressions are reported, just what is anyone supposed to conclude from them?

The notion of robustness here is an odd one, as can be seen from a simple

example. Let *A, B,* and *C* be mutually orthogonal variables. Let a linear combination of the three and a random error term determine a fourth variable *D*. Now if the coefficient on *C* relative to its variance is small compared with the coefficients on *A* and *B* relative to their variances and, the variance of *C* is small relative to the variance of the error term, then the coefficient on *C* may have a low conventionally calculated *t*-statistic and a high standard error. *C* has a low signal-to-noise ratio. Let us suppose that *C* is just significant at a conventional level of significance (say, five per cent) when the true specification is estimated. How will *C* fare under extreme bounds analysis? The omission of *A*, *B* or both, is likely to raise the standard error substantially and the point estimate of the coefficient on *C* plus or minus twice its standard deviation might now bracket zero.[7] We would then conclude that *C* is not a robust variable and that it is not possible to reach a consensus, even though *ex hypothesi* it is a true determinant of *D*.

One response might be that it is just an unfortunate fact that sometimes the data are not sufficiently discriminating. The lack of robustness of variable *C* tells us that, while there may be a truth, we just do not have enough information to narrow the range of prior beliefs about that truth, despite the willingness of investigators to consider the complete range of possibilities. The difference between the real world and the example here is that, unlike here we never *know* the actual truth. Thus, if we happen to estimate the truth, yet the truth is not robust, our true estimate carries little conviction or epistemic warrant.

A second response, however, is that the example here illustrates that there is no good reason to expect a true specification to be robust – that is, to be robust to mis-specification. Robustness is not, in general, a carrier of epistemic warrant. Leamer (in Hendry *et al.* 1990: 188) attacks the very notion of a true specification:

> I . . . don't think there is a true *data-generating process* . . .
>
> To me the essential difference between the Bayesian and a classical point of view is not that the parameters are treated as random variables, but rather that the sampling distributions are treated as subjective distributions or characterizations of states of mind . . . And by 'states of mind' what I mean is the opinion that it is useful for me to operate as if the data were generated in a certain way.

Econometrics for Leamer is about characterizing the data but not about discovering the actual processes that generated the data. We find this position to be barely coherent. The relationships among data are interesting only when they go beyond the particular factual context in which they are estimated. If we estimate a relationship between prices and quantities, for instance, we might wish to use it predictively (what is our best estimate of tomorrow's price?) or counterfactually (if the price had been different, how would the quantity have been different?). Either way, the relationship is meant to go

beyond the observed data and apply with some degree of generality to an unobserved domain. To say that there is a true data-generating process is to say that a specification could in principle, at least approximately, capture that implied general relationship. To deny this would appear to defeat the purpose of doing empirical economics. The very idea of a specification in which different observations are connected by a *common* description seems to imply generality. The idea of Bayesian updating of a prior with new information seems to presuppose that the old and the new information refer to a common relationship among the data – generality once more.

5 THE TRUTH IS SPECIALLY FITTED TO SURVIVE

The third attitude to data mining embraces the notion that there is a true data-generating process, although recognizing that we cannot ever be sure that we have uncovered it. A good specification-search methodology is one in which the truth is likely to emerge as the search continues on more and more data. On this view, data mining is not a term of abuse but a description of an essential empirical activity. The only issue is whether any particular data mining scheme is a good one. This pro-data mining attitude is most obvious in the so-called LSE (London School of Economics) methodology.[8] The relevant LSE methodology is the *general-to-specific modelling approach*. It relies on an intuitively appealing idea. A sufficiently complicated model can, in principle, describe the economic world.[9] Any more parsimonious model is an improvement on such a complicated model if it conveys *all* of the same information in a simpler, more compact form. Such a parsimonious model would necessarily be superior to all other models that are restrictions of the completely general model except, perhaps, to a class of models nested within the parsimonious model itself. The art of model specification in the LSE framework is to seek out models that are valid parsimonious restrictions of the completely general model and, that are not redundant in the sense of having an even more parsimonious model nested within them that also are valid restrictions of the completely general model.

The general-to-specific modelling approach is related to the theory of *encompassing*.[10] Roughly speaking, one model encompasses another if it conveys all of the information conveyed by another model. It is easy to understand the fundamental idea by considering two non-nested models of the same dependent variable. Which is better? Consider a more general model that uses the non-redundant union of the regressors of the two models. If model I is a valid restriction of the more general model (e.g., based on an F-test) and model II is not, then model I encompasses model II. If model II is a valid restriction and model I is not, then model II encompasses model I. In either case, we know everything about the joint model from one of the restricted models, we therefore know everything about the other restricted model from that one. There is, of course, no necessity that either model will be a valid

restriction of the joint model: each could convey information that the other failed to convey. A hierarchy of encompassing models arises naturally in a general-to-specific modeling exercise. A model is tentatively admissible on the LSE view if it is congruent with the data in the sense of being: (i) consistent with the measuring system (e.g., not permitting negative fitted values in cases in which the data are intrinsically positive); (ii) coherent with the data in that its errors are innovations that are white noise as well as a martingale difference sequence relative to the data considered; and (iii) stable (cf. Phillips 1988: 352–53; Mizon 1995: 115–22; White 1990: 370–74). Further conditions (e.g., consistency with economic theory, weak exogeneity of the regressors with respect to parameters of interest, orthogonality of decision variables) may also be required for economic interpretability or, to support policy interventions or other particular purposes. If a researcher begins with a tentatively admissible general model and pursues a chain of simplifications, at each step maintaining admissibility and checking whether the simplified model is a valid restriction of the more general model, then the simplified model will be a more parsimonious representation of all the models higher on that particular chain of simplification and will encompass all of the models lower along the same chain.

The general-to-specific approach might be seen as an example of invidious data mining. The encompassing relationships that arise so naturally apply only to a specific path of simplifications. One objection to the general-to-specific approach is that there is no automatic encompassing relationship between the final models of different researchers, who have wandered down different paths in the forest of models nested in the general model. One answer to this is that any two models can be tested for encompassing either through the application of non-nested hypothesis tests or through the approach described above, of nesting them within a joint model. Thus, the question of which, if either, encompasses the other can be resolved, except in cases in which sample size is inadequate.

A second objection notes that variables may be correlated either because there is a genuine relation between them or because – in short samples – they are adventitiously correlated. This is the objection of Hess *et al.* (1998) that the general-to-specific specification search of Baba *et al.* (1992) selects an 'overfitting' model. Any search algorithm that retains significant variables will be subject to this objection since adventitious correlations are frequently encountered in small samples. They can be eliminated only through an appeal to wider criteria, such as agreement with *a priori* theory. One is entitled to ask though, before accepting this criticism, on what basis should these criteria be privileged?

By far the most common reaction of critical commentators and referees to the general-to-specific approach questions the meaning of the test statistics associated with the final model. The idea of Darwinian test statistics arises, as it does for Lovell, because test statistics which are well-defined only under the

correct specification, are compared across competing (and, therefore, necessarily not all correct) specifications.

The general-to-specific approach is straight-forward regarding this issue. It accepts that choice among specifications is unavoidable, that an economic interpretation requires correct specification and that correct specification is not likely to be given *a priori*. The general-to-specific search treats and focuses on the relationship between the specification and the data, rather than, as is the case with the other two attitudes, on the relationship between the investigator and the specification. That is, it interprets the test statistics as evidence of sampling distribution rather than as measures of epistemic warrant. Each specification is taken on probation. The question posed is counterfactual: what *would* the sampling distributions be if the specification in hand were in fact the truth? The true specification, for example, by virtue of recapitulating the underlying data-generating process, should show errors that are white noise innovations. Similarly, the true specification should encompass any other specification (in particular it should encompass the higher dimensional *general* specification in which it is nested).

The general-to-specific approach is Darwinian but in a different sense than that implied in the other two attitudes. The notion that only our prejudices survive, or that the key issue is to modify critical values to account for the degree of search, assumes that we should track some aspect say, the coefficient on a particular variable, through a series of mutations (the alternative specifications) and that the survival criterion is our particular prior commitment to a value, sign or level of statistical significance for that variable. The general-to-specific methodology rejects the idea that it makes sense to track an aspect of an evolving specification. Since the specification is regarded as informative about the data rather than about the investigator or the history of the investigation, each specification must be evaluated independently. Nor should our preconceptions serve as a survival index. Each specification is evaluated for its verisimilitude (does it behave statistically like the truth would behave were we to know the truth?) and, for its relative informativeness (does it encompass alternative specifications?). The surviving specification in a search is a model of the statistical properties of the data and identical specifications bear the same relationship to the data whether that search was an arduous bit of data mining or a directly intuited step to the final specification.

Should we expect the distillation process to lead to the truth? The Darwinian nature of the general-to-specific search methodology can be explained with reference to a remarkable theorem due to Halbert White (1990: 379–80). The upshot of which is this: for a fixed set of specifications and a battery of specification tests, as the sample size grows toward infinity and increasingly smaller test sizes are employed, the test battery will – with a probability approaching unity – select the correct specification from the set. In such cases, White's theorem implies that type I and type II error both fall asymptotically to zero. White's theorem says that, given enough data, only the true specification will

survive a stringent enough set of tests. This turns the criticisms which regard data mining as Darwinian, in a pejorative sense, on their heads. The critics fear that the survivor of sequential tests survives accidentally and, therefore, that the critical values of such tests ought to be adjusted to reflect the likelihood of an accident. White's theorem suggests that the true specification survives precisely because the true specification is necessarily, in the long run, the fittest specification.

Approaches that focus on correcting critical values miss the point. White's theorem suggests that we envisage the problem differently. An analogy is the fitting together of a jigsaw puzzle. Even if a piece duplicates another in part or all of its shape, as more pieces are put into place, the requirement that the surface picture as well as the geometry of the pieces cohere implies that each piece has a unique position. Inferences about the puzzle as a whole, or the piece in relation to the puzzle, can be made soundly only conditional on getting the pieces into their proper positions. And, in the long run, the puzzle fits together only one way – a fact about the puzzle itself, not about us.

White's theorem is an asymptotic result. The real world of economics does not deal in infinite samples of data. But the general-to-specific methodology proceeds from a similar vision of the relationship of testing to the truth. The interesting methodological question on this view is, what are effective procedures for solving the jigsaw puzzles of economics when samples are small? We have made a first pass at this question.

Hoover and Perez (1999) evaluate the general-to-specific methodology in a simulation study inspired by Lovell's (1983) Monte Carlo study of mechanized search algorithms (see section 3 above). Lovell concludes that the three simple algorithms he examines (step-wise regression, maximum R^2 and max-min t-statistics) are all quite poor in recovering the true data-generating processes. Furthermore, the size and power of the algorithms taken as a whole (that is the ability of the algorithms to exclude variables that were not in the data-generating process and to include variables that were) is rather poor and quite different from that implied by conventional critical values based on sampling distributions. Using updated data and the same data-generating processes, we are able to confirm Lovell's results for the algorithms he tests on annual data.

We then extend the investigation to the general-to-specific search procedure. We use the same variables but at a quarterly frequency and we difference each series until it is stationary on standard tests. In the hands of econometricians of the LSE school, the general-to-specific methodology is not mechanical like Lovell's step-wise regression or other algorithms. Nevertheless, we have developed a mechanical algorithm that mimics some key features of the general-to-specific approach. It begins with a general model in which the entire set of variables in Lovell's data set (included lagged variables) appears as regressors. This regression is tested for congruence. If it passes, simplification begins. Regressors with the low t-statistics are deleted in sequence,

starting a different search with each of the ten lowest, providing that they are insignificant at conventional sizes. When a regressor is deleted and a new estimate obtained, it is checked to see whether it encompasses the general model. If it does not or if it fails any tests of congruence, the regressor is replaced and the variable with the next lowest t-statistic from the previous regression is eliminated instead. If it does, then the variable with the next lowest insignificant t-statistic in the current regression is eliminated and, the process of testing is repeated. Elimination continues until either all retained variables are significant, no variables can be removed without failing congruence, or no variables can be removed without failing to encompass the general model. Ten such regressions from ten search paths, corresponding to the ten regressors with lowest t-statistics in the general model, constitute the possible characterizations of the data. The model among these ten that encompasses the others is chosen as the final model.[11]

Lovell, and proponents of the first view of data-mining would have us report this specification but, adjust the standard errors of the coefficients to account for the many regressions run. Leamer and proponents of the second view of data mining would have us look at all of the many regressions and choose only those variables that do not have their coefficient values change significantly over the search, i.e. the end result is not important, only the distribution of the coefficients estimates is important. The general-to-specific approach takes the final specification to represent the best approximation to the truth because it acts as closely to how the truth would act, if we knew it. Evolution is important but not the history of evolution.

To check the success of this algorithm, we simulate with the updated quarterly data the same models that Lovell uses. There are nine models, static and dynamic, calibrated from the actual data. The independent variables are actual data but the dependent variable is simulated from the actual independent variables and draws from a random distribution with characteristics that match the performance of each model in actual data. We conduct 1000 simulations and searches of each model. In contrast to Lovell's algorithms, we find that the general-to-specific methodology is effective at recovering the true data-generating process. It does not always succeed. Where it fails, it seems to be almost exclusively because of low signal-to-noise ratios. This is reflected in the fact that both size and power measures are close to what one obtains from Monte Carlo simulations in which the true specification is known in advance. It also shows that no method could be expected in some cases to find the true model when there is simply not enough information in the data. Our results are supportive of the general-to-specific methodology but they are limited. Work-in-progress aims to extend the evaluation to non-stationary data, in which questions of cointegration are important, and to cross-sectional contexts.

The point of this long digression is not principally to advertise our own work although we are happy if it does that. Rather it is to demonstrate an empirical spirit. We hope to have provided a theoretical analysis of why the

concerns of the opponents of data mining are misplaced but we also wish to allay any nagging doubts by pointing to evidence that, practically, the difficulties they foresee do not in fact arise.

6 CONCLUSION

Econometrics is a tool for learning about the economy. It presupposes the existence of an economy to learn about and not an assortment of facts but, an economy with real general features that imply the behaviour of the data and constrain the way that econometric and statistical calculations package the data. There is a truth about the economy that remains the truth, even though it presents a different aspect when viewed through different econometric filters, just as there is a truth about the moon or the crab nebula that remains the truth even though these heavenly bodies *appear* differently when viewed through different telescopes and optical filters. The central issue is how to use these observations most effectively. The school of thought that argues that the more the search, the lower the epistemic warrant would reject a detailed picture of a distant galaxy because it was difficult to obtain. It concentrates on the astronomer, rather than on the astronomical object. The problem is that the object may be objectively difficult to find and may yield only to highly structured search. Indeed, many scientific results are valued precisely because they are difficult to obtain. The school of thought that seeks robustness rather than truth would reject the picture because many other pictures of the same sector of the sky processed in different ways look quite different. Whereas the school of thought that we endorse argues that the issues that need to be addressed are, first, whether the lenses and filters of the astronomers or the statistics of the econometricians are in fact the ones that *would* reveal the aspect of the truth that interests us – *if it were there* – and, second, whether, in the event, they do so. A regulated specification search, such as the general-to-specific methodology proposes, is an attempt to use econometrics to bring an economic reality into focus that would otherwise remain hidden. It aims, quite literally, to discover the truth.

Kevin D. Hoover
Department of Economics, University of California, USA
kdhoover@ucdavis.edu

Stephen J. Perez
Department of Economics, Washington State University, USA
sjperez@wsu.edu

ACKNOWLEDGEMENTS

The authors wish to thank Roger Backhouse, Peter Burridge, Alistair Hall, Thomas Mayer, Mary Morgan and Steven Sheffrin for comments on earlier drafts.

NOTES

1 This is not to deny that analytical conclusions are drawn with respect to models with stochastic regressors. But even such models are characterized at some level by probability distributions with constant parameters (cf. the debate between Leamer and Hendry in Hendry *et al.* (1990: 195).
2 To keep the discussion simple, we will often speak of the standard errors and *t*-statistics of regression coefficients as exemplars of the sampling distribution. Our points are also relevant *mutatis mutandis* to other statistics.
3 On Monte Carlo methods, see Hendry (1995, ch. 3, section 6); on bootstrap methods, see Jeong and Maddala (1993).
4 Sampling distribution and epistemic warrant are two distinct things. In particular, epistemic warrant refers to the circumstances in which we have adequate justification for a belief. It is not the perfection of the sampling distribution in the sense of being the population distribution that corresponds to the sample.
5 These categories are drawn from McAleer *et al.* (1985) and their restatement of Leamer's extreme bounds analysis.
6 It is important not to confuse the idea of robustness here with notions of temporal stability or homogeneity among subsamples. Truth need not be constant over time nor global for a given set of data. Robustness to alternative specification holds the sample constant, so these questions do not arise; it is only the specification that varies.
7 The point estimate itself will not switch signs if the regressors are orthogonal as supposed but, might do so if they are correlated with each other.
8 The LSE approach is described sympathetically in Gilbert (1986), Hendry (1987, 1995, especially ch. 9–15), Pagan (1987), Phillips (1988), Ericsson *et al.* (1990) and Mizon (1995). For more sceptical accounts, see Faust and Whiteman (1995) and Hansen (1996). The adjective 'LSE' is, to some extent, a misnomer. It derives from the fact that there is a tradition of time-series econometrics that began in the 1960s at the London School of Economics, see Mizon (1995) for a brief history. The practitioners of LSE econometrics are now widely dispersed among academic institutions throughout Britain and the world.
9 This is a truism. Practically, however, it involves a leap of faith; for models that are one-to-one, or even distantly approach one-to-one, with the world are not tractable.
10 For general discussions of encompassing, see, for example, Mizon (1984), Mizon and Richard (1986), Hendry and Richard (1987), Hendry (1988, 1995 ch. 14).
11 In this paper we use the necessary condition for encompassing that the standard error of regression for the encompassing model must be the lowest of the ten. In work in progress we conduct a sequence of encompassing tests to choose among the ten, allowing for the fact that some linear combination of them might in fact encompass the others, even while none of them does so individually.

REFERENCES

Baba, Yoshihisa, Hendry, David F. and Starr, Ross M. (1992) 'The demand for M1 in the USA', *Review of Economic Studies* 59: 25–61.

Ericsson, Neil R., Campos, Julia and Tran, H.-A. (1990) 'PC-GIVE and David Hendry's econometric methodology', *Revista de Econometrica* 10: 7–117.

Faust, Jon and Whiteman, Charles H. (1995) Commentary [on Grayham E. Mizon 'Progressive modeling of macroeconomic times series: the LSE methodology'], in Kevin D. Hoover (ed.) *Macroeconometrics: Developments, Tensions and Prospects*, Boston: Kluwer, pp. 171–80.

Gilbert, Christopher L. (1986) 'Professor Hendry's econometric methodology', *Oxford Bulletin of Economics and Statistics* 48: 283–307. (Reprinted in Granger 1990).
Goldfeld, Stephen M. (1973) 'The demand for money revisited', *Brookings Papers on Economic Activity*, no. 3: 577–638.
Granger, C.W.J. (1990) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford: Clarendon Press.
Hansen, Bruce E. (1996) 'Methodology: alchemy or science?', *Economic Journal* 106: 1398–431.
Hendry, David F. (1987) 'Econometric methodology: a personal viewpoint', in Truman Bewley (ed.) *Advances in Econometrics*, Vol. 2, Cambridge: Cambridge University Press.
Hendry, David F. (1988) 'Encompassing', *National Institute Economic Review* 125: 88–92.
Hendry, David F. (1995) *Dynamic Econometrics*, Oxford: Oxford University Press.
Hendry, David F., Leamer, Edward E. and Poirier, Dale J. (1990) 'The ET dialogue: a conversation on econometric methodology', *Econometric Theory* 6: 171–261.
Hendry, David F. and Richard, Jean-François (1987) 'Recent Developments in the Theory of Encompassing', in Bernard Cornet and Henry Tulkens (eds) *Contributions to Operations Research and Economics: The Twentieth Anniversary of Core*, Cambridge, MA: MIT Press, pp. 393–440.
Hess, Gregory D., Jones, Christopher S. and Porter, Richard D. (1994) 'The predictive failure of the Baba, Hendry and Starr model of the demand for M1, *Journal of Economics and Business* 50: 477–507.
Hoover, Kevin D. (1995) 'In defense of data mining: some preliminary thoughts', in Kevin D. Hoover and Steven M. Sheffrin (eds) *Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer*. Aldershot: Edward Elgar.
Hoover, Kevin D.and Perez, Stephen J. (1999) 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search', *Econometrics Journal* 2: 1–25.
Jeong, J. and Maddala, G.S. (1993) 'A perspective on applications of bootstrap methods in econometrics', in G.S. Maddala, C.R. Rao and H.D. Vinod (eds) *Handbook of Statistics*, vol. 11, *Econometrics*, Amsterdam: North Holland, pp. 573–610.
Judd, John P. and Scadding, John L. (1982) 'The search for a stable money demand function: a survey of the post-1973 literature', *Journal of Economic Literature* 22: 993–1023.
Leamer, Edward (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Boston: John Wiley.
Leamer, Edward, (1983) 'Let's Take the Con Out of Econometrics', *American Economic Review* 73: 31–43. Reprinted in Granger (1990).
Leamer, Edward E. and Leonard, Herman (1983) 'Reporting the Fragility of Regression Estimates', *Review of Economics and Statistics* 65: 306–317.
Lovell, Michael C. (1983) 'Data mining', *The Review of Economics and Statistics*, 65: 1–12.
Mayer, Thomas (1980) 'Economics as a Hard Science: Realistic Goal or Wishful Thinking?', *Economic Inquiry*, 18: 165–78.
Mayer, Thomas (1993) *Truth versus Precision in Economics*. Aldershot: Edward Elgar.
Mayer, Thomas (2000) 'Data mining: a reconsidertion', *Journal of Economic Methodology* 7: 183–90.
McAleer, Michael, Pagan, Adrian R. and Volker, Paul A. (1983) 'What will take the con out of econometrics', *American Economic Review* 75 (3) June: 293–307. Reprinted in Granger (1990).
Mizon, Grayham E. (1984) 'The Encompassing Approach in Econometrics', in

D.F. Hendry and K.F. Wallis (eds) *Econometrics and Quantitative Economics,* Oxford: Basil Blackwell, pp. 135–72.

Mizon, Grayham E. (1995) 'Progressive modelling of macroeconomic time series: the LSE methodology', in Kevin D. Hoover (ed.) *Macroeconometrics: Developments, Tensions and Prospects*, Boston: Kluwer, pp. 107–70.

Mizon, Grayham E. and Richard, Jean-François (1986) 'The encompassing principle and its application to testing non-nested hypotheses', *Econometrica* 54, 657–78.

Pagan, Adrian (1987) 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys* 1: 3–24. Reprinted in Granger (1990).

Phillips, Peter C. B. (1988) 'Reflections on econometric methodology', *Economic Record* 64: 334–59.

White, Halbert (1990) 'A Consistent Model Selection Procedure Based on *m*-testing', in C.W.J. Granger (ed) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford: Clarendon Press, pp. 369–83.