The miracle of the Septuagint and the promise of data mining in economics

STAN DU PLESSIS

Stellenbosch Economic Working Papers: 15/06

KEYWORDS: Data mining, model selection, automated model selection, general to specific modelling, extreme bounds analysis, Bayesian model selection
JEL: C110, C500, C510, C520, C870

STAN DU PLESSIS
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: STAN@SUN.AC.ZA

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

The miracle of the Septuagint and the promise of data mining in economics

STAN DU PLESSIS[1]

ABSTRACT

This paper argues that the sometimes-conflicting results of a modern revisionist literature on data mining in econometrics reflect different approaches to solving the central problem of model uncertainty in a science of non-experimental data. The literature has entered an exciting phase with theoretical development, methodological reflection, considerable technological strides on the computing front and interesting empirical applications providing momentum for this branch of econometrics. The organising principle for this discussion of data mining is a philosophical spectrum that sorts the various econometric traditions according to their epistemological assumptions (about the underlying data-generating-process DGP) starting with nihilism at one end and reaching claims of encompassing the DGP at the other end; call it the DGP-spectrum. In the course of exploring this spectrum the reader will encounter various Bayesian, specific-to-general (S-G) as well general-to-specific (G-S) methods. To set the stage for this exploration the paper starts with a description of data mining, its potential risks and a short section on potential institutional safeguards to these problems.

Keywords: Data mining, model selection, automated model selection, general to specific modelling, extreme bounds analysis, Bayesian model selection
JEL codes: C110, C500, C510, C520, C870

[1] Written for the Oxford Handbook of Philosophy and Economics and first presented at the University of Alabama in Birmingham during May 2006. I am grateful for the many helpful comments from the conference participants and for specific comments by Ben Smit. The usual disclaimer applies.

The miracle of the Septuagint and the promise of data mining in economics

WHILE hard pressed in his second round with Tinbergen, Keynes played a trump: the Septuagint, he reminded Tinbergen, was produced by seventy translators working independently from the same Hebrew text and who emerged from their cells to find, miraculously, that they had produced seventy identical Greek translations of the Old Testament. And so the Septuagint was held to carry the (considerable) authority of independent confirmation. In modern econometric parlance the Septuagint might be called a robust translation – a believable text for the Greek world – or it might simply have been the truth.

Turning on Tinbergen, Keynes now wondered: "Would the same miracle be vouchsafed if seventy multiple correlators were shut up with the same statistical material?" and how else might we distinguish econometrics from "statistical alchemy" (Keynes, 1940: 155-156)? It is a question that reveals unease with data mining: that rival and inconsistent models might proliferate, that design criteria might reveal the prejudices of the modeller not the underlying economic structure, that parameter estimates might be biased and that test sizes might be misleading. In short, data mining could compromise the believability of econometric models.

And yet there might not be an alternative to data mining if economics is going to be an empirical science practiced with the joint constraints of incomplete economic theory and non-experimental data. We need data to complement our otherwise inadequate theoretical models, but because economic data is only rarely experimental, that means repeated use of the same data. This leaves modern econometricians to steer, as David Hendry (1997) said, between the Scylla of theory-dependence and the Charybdis of data-dependence[2]. Doing so with data mining seems to offend against norms of good conduct in econometrics; or at least, such norms were widely shared until recent advances in the theory of econometric modelling (Backhouse and Morgan, 2000) and access to unprecedented computing capacity[3] had encouraged a revisionist literature in which the unhappy connotations of data mining have yielded to the view that data mining is a necessary part of a sensible modelling strategy (Granger, 1999; Hendry and Krolzig, 1999; Hoover and Perez, 1999; Phillips, 2005). Indeed, data mining may even be a virtue (Greene, 2000; Hoover and Perez, 2000). This paper considers these developments and their implications for data mining in econometrics.

The organising principle for this discussion of data mining is a philosophical spectrum that sorts the various econometric traditions according to their epistemological assumptions (about the underlying data-generating-process DGP) starting with nihilism at one and reaching claims of encompassing the DGP at the other end; call it the DGP-spectrum. In the course of exploring this spectrum the reader will encounter various Bayesian, specific-to-general (S-G) as well general-to-specific (G-S) methods. To set the stage for this exploration the paper starts with a description of data mining and its potential dangers and a short section on potential institutional

[2] Pagan (2003) uses the same tension between data-dependence and theory-dependence as an organising principle for his survey of modern macroeconometric modelling methods.

[3] The dramatic potential of modern computing power is transforming research in many fields (Glymour, 2004), and there is optimism that it will do the same for applied econometrics (Phillips, 2005).

safeguards to these problems.

1. Shared experience

Chris Chatfield uses the term "data mining" to describe the situation where "… models are not fully specified *a priori*, but rather are formulated, at least partially, by looking at the *same* data as those later used to fit the model" (Chatfield, 1995: 426, emphasis in the original). This definition highlights the same features emphasised by Hoover and Perez (2000), that is: the use of data to describe and estimate models on the one hand and the use of the same data to evaluate models against certain design criteria.

Both definitions emphasise the *dual use of data* and this runs like a thread through the various habits associated with data mining in econometric modelling (Chatfield, 1995; Mayo, 1996; White, 2000). Such dual use of data could manifest in more than one way, including (Spanos, 2000): selecting a data set or sample, choosing regressors, the respecification of a model, diagnostic testing and visual inspection of the data (data snooping).

The type data mining at stake in this paper is this approach to modelling characterised by the dual use of data, not that branch of statistics, also called "data mining", that studies the use of modern computer algorithms to search for patterns in large databases. Such statistical data mining is not an attempt to uncover any underlying DGP (Hand, 1998). The boundary between this statistical study of local patterns in data and data mining as a tool of econometric modelling is not strict though, indeed econometrics at the nihilistic end of the spectrum introduced above overlaps to some extent with the statistical tradition of data mining.

Data mining, at any point along the DGP spectrum, implies the repeated use of the same data. However, since so much economic data is non-experimental (also called observational data) the repeated use of economic data is difficult, or even impossible, to avoid (Spanos, 1995; White, 2000). "Econometrics", so Schumpeter argued early on, "is nothing but the explicit recognition of this rather obvious fact" (Schumpeter, 1933: 6).

The statistical considerations relevant to econometrics with non-experimental data are different from those of experimental statistics (Spanos, 1995; Hand, 1998). Indeed, Spanos has argued that given the inevitably widespread use of non-experimental data in economics[4], it would be more sensible to locate econometrics in the biometric tradition of statistics than in the experimental design tradition (Spanos, 1995; 1999). An econometrician operating in the latter tradition respects what Spanos calls a *predesignationist* rule that requires a specification of the relevant hypothesis before examining the data (Spanos, 2000). In contrast the biometric tradition was explicitly developed for settings where such rules are irrelevant.

[4] Schumpeter's well-known claim for economics as the "most quantitative" of all sciences is based on the non-experimental nature of much economic data. In his words from the first edition of Econometrica "There is, however, one sense in which economics is the most quantitative, not only of 'social' or 'moral' sciences, but of *all* sciences, physics not excluded. For mass, velocity current, and the like *can* undoubtedly be measured, but in order to do so we must always invent a distinct process of measurement…Some of the most fundamental economic facts, on the contrary, already present themselves to our observation as quantities made numerical by life itself. They carry meaning only by virtue of their numerical character…Econometrics is nothing but the explicit recognition of this rather obvious fact, and the attempt to face the consequences of it" (Schumpeter, 1933: 5-6, emphasis in the original).

The observational nature of much economic data implies an unknown sampling model and this adds to model uncertainty born from incomplete theoretical models. Such model uncertainty manifests in a number of ways: the variables to be included, the functional and probabilistic form of the model and the choice between rival models (Chatfield, 1995). This challenge is particularly acute with small samples (Pagan and Veall, 2000), since White's theorem guarantees the recovery of the true DGP from an over-parameterised initial model in large samples[5] (White, 2005 [1990]).

There is a widely held perception that data mining in the sense of a dual use of non-experimental data in the context of model uncertainty is widespread (Chatfield, 1995; Greene, 2000; Mayer, 2000; Pagan and Veall, 2000; White, 2000). Burger and du Plessis (2006) have quantified the extent of data mining in applied econometrics by evaluating the modelling strategies employed in a random sample of papers drawn from academic journals in economics published in 2003. From the 75 papers in their sample 71% used non-experimental data and 89% showed explicit evidence of an iterative modelling strategy, of which the dual use of data associated with data mining is a common form.

For the remainder of this essay this combination of model uncertainty and observational data, which is associated with data mining in practice, will be treated as a shared experience for applied econometrics. While all the traditions considered here share this experience, they move beyond it with the guidance of different theories and techniques and in opposite directions.

2. The risks of data mining

The collection of research habits gathered under the heading of data mining has frequently furnished a stick with which to beat the econometric fraternity, and to cast doubt over the value of the considerable annual econometric output (for example, Leamer, 1978; Karni and Shapiro, 1980; Leamer, 1983; Lovell, 1983; Denton, 1985). It has also been used by proponents of specific econometric methods in their criticism of rival methods: Hendry (1997), for example, argued that Faust and Whiteman (1997a) had underplayed the "incipient 'data mining'" in RBC (Real Business Cycle) and VAR (vector autoregression) methods, while Faust and Whiteman retorted that "…every LSE-style paper reports extensive theory-free tailoring of the model that overwhelms any such alterations made in the RBC literature" (Faust and Whiteman, 1997b: 192).

A recent example by Sullivan, Timmerman and White (2001) show that these arguments are not just rhetorical. Instead, they are based on the real risk that data mining might undermine the statistical inference at the heart of applied econometrics. In an investigation of calendar effects on the stock market Sullivan et al. (2001) used a

[5] This result is subject to the conditions mentioned in the discussion on the G-S method below. Absent these conditions Rissanen's theorem may be applied to show that the DGP will not be recovered, even asymptotically (Phillips, 2003).

technique developed by White (2000) to evaluate the predictive power of different calendar-based trading rules, while accounting for the large number of potential rules amongst which the candidate rules would be selected. Their results were startling: while a number of individual calendar-effects were significant, none of these remained so once the extent of data mining that had accompanied their discovery was factored in using White's "reality check".

Stepping away from examples, the various risks posed by data mining could be divided into the following categories (Spanos, 2000)[6]: (i) data and sample selection; (ii) selection of regressors; (iii) respecification, and (iv) diagnostic testing. But these are only risks, and a framework is needed to understand when the risk is likely to lead to undesirable outcomes.

To that end Spanos (2000) built on the econometric and philosophy of science literatures to derive a definition of "unwarranted" data mining. The latter occurs when a researcher (i) interprets data as evidence in support of a proposition (or theory or hypothesis), having (ii) searched either over the data to establish such evidence or having searched for data supportive of the proposition, and while (iii) the proposition would fail a severe test on the same data. It is the combination of all three these aspects that causes mischief. This is the definition of unwarranted data mining used throughout this paper.

The specification searches (within and across data sets) mentioned by Spanos (2000) does not itself present a insurmountable problem to the careful econometrician (Spanos, 2000: 235). The various traditions discussed in sections 4 through 6 are so many different ways of using such searches without – so their proponents claim – incurring the risks of unwarranted data mining. These traditions are often proponents of specific modelling strategies, and this is not by accident: a consequence of model uncertainty is that the modeller has no choice but to actively engage in model specification, which entails: formulating the model, estimating relevant parameters and evaluating the model all of which occurs in an "iterative and interactive way" (Chatfield, 1995: 425). But these "iterative and interactive" steps have to taken with a watchful eye on the risks that are briefly explored in the following few paragraphs.

2.1 Searching for regressors

The shared experience behind this discussion of data mining is the combination of model uncertainty and non-experimental data. The risks posed by "searching for regressors" are closely related to model uncertainty while the non-experimental nature of the data moves to the fore in the discussion of "data and sample selection" below. Due to model uncertainty, econometricians usually have considerable leeway in the choice of variables, the combinations of variables to be included and the functional form of any estimable model.

The risk of unwarranted data mining looms large in this case, especially when the following iterative strategy is followed: select combinations of potential explanatory factors iteratively until the coefficients of the model are statistically significant at a conventional level (such as 5%). In what has become a famous paper Lovell (1983) used a Monte Carlo simulation to show that the probability of type I errors are much larger in such an iterative

[6] For alternative discussions along the same lines see Leamer (1978) and Chatfield (1995).

process than the conventional 1% or 5% level of the final test.

For example, in the simple case where an econometrician had been searching for the "best" two regressors out of ten potential regressors – not an unusually large number for the empirical growth literature, e.g. Sala-i-Martin (1997) – in a simulation where the true DGP had precisely two regressors, Lovell found that the true significance level of a t-test with 5% nominal size was, in his simulation, a much higher 22.6% (Lovell, 1983: table 1).

In this example the statistical tests conducted after a dual use of the data were no longer "severe" tests. There is a long literature in economics that explores estimation bias and implied changes to the size of tests from similar dual use of the data (Wallace and Ashar, 1972; Denton, 1985; Giles and Giles, 1993; Chatfield, 1995; Granger et al., 1995; Spanos, 1995). Perhaps Leamer has articulated their collective concern most forcefully:

> "This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation, in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose" (Leamer, 1983: 36-37).

The various biases caused by data mining increases the likelihood that the final specification will show spurious relationships, instead of uncovering true effects. This fear of spurious correlation is a prime reason for the professional suspicion of data mining in econometrics (Hoover and Perez, 2000).

In response to this risk one might follow Lovell's (1983) early recommendation that authors be required to state explicitly the extent of the search they conducted and to adjust the sizes of tests accordingly. Statisticians like Chatfield (1995) have argued along similar lines that the statistical report on parameter estimates remains incomplete until the model selection strategy is taken into account.

This recommendation is elaborated on below (in section 4) but as an empirical matter there is little evidence that it has had much impact on applied econometrics. Of the 67 papers that showed evidence of an iterative modelling strategy in Burger and du Plessis (2006), just three papers gave an indication of the extent of iterations involved and only one of these allowed the extent of the iterations to influence the statistical inference. Further, the data mining of any one researcher working in a field of researchers represents only a fraction of the total number of searches[7] (Denton, 1985). In such circumstances the appropriate adjustment to test sizes requires information about search paths in the entire literature. Uncovering such information is likely to run into severe difficulties given, *inter alia*, the observed reluctance of authors to publish an indication of their model searches (mentioned above).

[7] For Denton (1985) it is a fallacy of composition to argue that the risks of data mining would be diminished if everybody avoided data mining in their own research.

2.2 Data and sample selection

A second variant of data mining is the iterative search for a data set. Such a search risks becoming unwarranted data mining when the modeller is willing to search over alternative data sets (differing in: sample period; or nature, i.e. panel, cross-section or time series; frequency; or even different empirical measures of the same theoretical construct[8]) until the data yields an estimated model consistent with her theoretical priors.

This variant of data mining is unwarranted when the final model is offered as support for a theoretical proposition after an iterative data search that constructed in order to confirm the theoretical proposition. The probability of a type II error is large in this case, as an estimated model contradicting the theoretical proposition is easily interpreted as evidence against the data, rather than evidence against the theory (Spanos, 2000).

The use of "new data" is often mentioned as a potential safeguard against this variant of data mining (as it also believed to be against other variants of data mining) (for example, Granger, 1999; Greene, 2000). "New data" might mean a new sample of the same population, or cross validation in cross-section data or post-estimation (hold-out) samples in time series or a pure *ex ante* forecast. The hope is that new data would offer severe tests of the theoretical proposition at stake by introducing an element of objectivity in econometric modelling[9].

It is especially for this reason that an *ex ante* forecast is sometimes thought to be the "gold standard" of model evaluation (Clements and Hendry, 2005). However, Clements and Hendry (2005) have recently argued that this exaggerates the role that forecast performance can play in econometric the model evaluation (see also, O'Hagen, 1995; Sugden, 1995). Their reasoning is based on eight dichotomies which, when taken together, undermines the view that forecast performance is a (or *the*) key test in model evaluation.

The dichotomies are: unconditional versus conditional models; internal versus external standards for forecast evaluation; checking constancy versus adventitious significance; *ex ante* versus *ex post* evaluation; one-step versus multi-step forecasts; fixed coefficients versus updating; stationarity versus non-stationarity and, finally, forecasting versus other objectives (Clements and Hendry, 2005). While this is not the place to discuss the details of these dichotomies here their collective impact is to undermine the (sometimes exaggerated) role given to a specific forecast performance in the process of model evaluation. An economist has to evaluate her project carefully with respect to all eight of these dichotomies before making strong claims for the ability of an ex ante forecast to judge an econometric model.

It is possible to add an empirical observation to Clements and Hendry's (2005) theoretical caution, that is: new

[8] For example, searching over the various proxies for "institutional quality" to identify an institutional effect in a growth regression.

[9] The objectivity of econometrics, as with other scientific activities depends *inter alia* on the inter-subjective testability of the models. All econometricians, like all other scientists are individually subjective in their treatment of theories and data, and the objectivity of the scientific exercise emerges, if at all, through the inter-subjective testing of these models by other econometricians and especially on new data (Popper, 1992 [1961]; 2000 [1959]). The physicists Feynman (1998: 18) acknowledges a "number of special techniques associated with the game of making observations" not unfamiliar to economists concerned with data mining, and then he come to the same conclusion as Popper (2000 [1959]), i.e. that the objectivity of data is only to be found in repeated testing, and not only by yourself.

data is but little used to safeguard econometric models from unwarranted data mining. Burger and du Plessis (2006) found only 6 papers that used "new data" as a part of the model evaluation amongst the 75 papers examined by them.

2.3 Respecification

Unwarranted data mining might also occur for a third reason, i.e. the *ad hoc* adjustment of, say, the functional form of models in response to unfavourable output from statistical tests. In this way the modeller uses the unfavourable statistical output inductively to respecify the model in such a way that, given the data, the same test is passed by the respecified model. It follows that the probability of passing this test with the respecified model is very high. But, by the same token the severity of the test would have been greatly reduced (Spanos, 2000).

Popper (2000 [1959]) had earlier argued that such *ad hoc* adjustments reduced the "empirical content" of a theory, where he defined the latter as the "class of potential falsifiers". He argued that any hypothesis could always be "immunised" against falsification by finite observations through the use of *ad hoc* auxiliary hypotheses. Popper proposed a methodological rule to avoid the unwarranted data mining implicit in such a strategy, i.e. that auxiliary hypotheses should increase, not decrease, the testability of the theory in question (Popper, 2000 [1959]).

The dynamic specification of time series models has frequently served as a demonstration of this risk in econometrics. An estimated macroeconomic time series model often shows evidence of residual autocorrelation, perhaps detected by the Durbin-Watson statistic. It is an easy step – though a *non-sequitur* – from this evidence to the conclusion that the stochastic error of the underlying DGP has an autocorrelated structure. A respecified model that "corrects" for this suspected autocorrelated error structure "solves" the problem of autocorrelated residuals, even when no such structure exists for the DGP.

This is the danger of unwarranted data mining by respecification: the modeller shoots the messenger by eliminating the information content of statistical tests that were meant to warn her that the estimated model conflicted with the data (Hendry, 1980; Spanos, 1986; Hendry, 1995). Elsewhere, Hendry expressed the danger as follows:

> "[*ad hoc* respecification] ensures that it [the model] matches where it touches – but otherwise leads to invalid inference…a revision process of gradually expanding a model and stopping at the first insignificant improvement maximises the initial contamination and hence the likelihood of false inferences" (Hendry, 2000 [1985]: 275) .

2.4 Diagnostic testing

Unwarranted data mining compromises not just estimation, but model evaluation too. Due to model uncertainty and the subsequent dual use of data, econometrics requires a discipline of post-estimation model evaluation. But this discipline is poorly developed, as Clive Granger (1999) has lamented, with applied econometrics often reflecting greater concern with model inputs (for example, the estimation procedure) than with model outputs (for example, forecasts on new data).

Diagnostic testing is only infrequently an important criterion in model selection (Krämer et al., 1985)[10], and Kennedy (2003) has offered some reasons for the observed scepticism towards diagnostic testing, including: that their validity depends on the validity of the estimated model and that a series of such tests affects the sizes of these tests. These problems are not insurmountable and the G-S approach (discussed below) has a long literature showing that the conditions required for informative diagnostic testing are crucial elements in a progressive modelling strategy[11].

In short G-S authors allow that the validity of diagnostic tests depend on the validity of the estimated model. To establish the latter we need to draw a distinction between testing primary hypotheses on the one hand and mis-specification tests, on the other. The latter are Fisher tests, i.e. tests "'without' the boundaries of the postulated model", while primary hypotheses are estimated "within" the bounds of a postulated model (Spanos, 2000: 257).

The extensive use of the mis-specification tests is a critical part of the G-S strategy in the service of the explicit goal of finding a statistically adequate general model[12]. However, the battery of misspecification tests is always a once-off occurrence for a newly proposed general model, so that there is no risk of reduced test size in establishing the statistical adequacy of the general model (Campos et al., 2005b). The G-S approach, with its emphasis on the destructive role of misspecification tests avoids unwarranted data mining from such tests, but this result depends critically on the G-S goal of locating a statistically adequate model, which is congruent with the local DGP (Hendry, 1995)[13]. The other formal responses to data mining discussed below do not share this goal, and are, as a consequence, more circumspect about the potential safeguard from unwarranted data mining offered by these tests.

In summary, unwarranted data mining risks undermining applied econometrics in a number of ways: it risks compromising the foundations of the associated statistical inference especially by reducing the severity of tests. In their recent reflection on data mining Hoover and Perez (2000) suggested three possible attitudes towards the reality of data mining: first, don't do it, but if you must, then adjust the tests accordingly. Second, you can't avoid data mining, hence you should investigate and (somehow) report al possible model specifications. Thirdly, data mining is an essential part of reasonable econometrics, but needs to be implemented in the right way.

From the implicit data mining present in any examination of existing theories or existing data – (Denton, 1985; Greene, 2000) – and the empirical investigation in Burger and du Plessis (2006) it is evident that the first response considered by Hoover and Perez (2000) is little practiced. This leaves the second and third attitude to explore, but before turning to the various methodological responses entailed by those two attitudes, we take a short detour to

[10] Krämer et al. (1985) found that an overwhelming proportion of published models in their sample failed standard diagnostic tests.
[11] Hendry defines a "progressive modelling strategy" as one where successive models account at least for the information in existing models, i.e. the new model encompasses the existing models (Hendry, 2000 [1985]).
[12] See, for example, Hendry's three golden rules for econometrics "test, test and test" (Hendry, 1980).
[13] The modeller only proceeds with the testing of primary hypotheses once the statistical adequacy of the general model has been established in the G-S approach. Such hypotheses take the form of restrictions on the general model and the inference is supported by the statistical adequacy of the general model (Spanos, 1999).

consider possible institutional features, incentives and disincentives, that might affect the risk of unwarranted data mining.

3. Institutional considerations

The catalogue of risks associated with unwarranted data mining adds force to normative campaigns for appropriate reform to econometric method and practice. Nevertheless, and as Pagan and Veall (2000) recently observed, data mining remains ubiquitous despite decades of unease. Such persistence in the face of moral disapproval might encourage economists to investigate the positive causes of widespread data mining.

Following Pagan and Veall (2000)one could either argue that economists have revealed a preference for data mining or that data mining is a market outcome in the decentralised market for scholarly research and publication (Pagan and Veall, 2000). Both interpretations open the way to an analysis of the incentive structure that supports data mining; that is, consideration of the formal and information institutions of applied econometric research.

While such an analysis is beyond the scope of this paper, the role that these institutions could play as a safeguard against unwarranted data mining is not. Editors, as intermediaries between the producers and consumers of econometrics, have the opportunity to impose standards that might facilitate the competitive evaluation of papers (Pagan and Veall, 2000). To the extent that data is made available to referees, and the journal's audience, the direct feedback effect and the indirect feedback effect through reputation effect could contribute strengthen safeguards against unwarranted data mining. Important results (especially with policy implications) might eventually be subjected to further tests, possibly by rival economists. In this way the passage of time (which creates new un-mined data) and the participation by rival researchers is likely to uncover fragile results.

The incentives created by editorial policies could also counter or nurture such critical evaluation of published work. For example, Denton (1985) warned of a "publication filter" whereby editors look more favourably on papers that report "significant" econometric results. Such a filter does not need formal or even informal enforcement by editors, as self-selection by potential authors who mine their models until they are able to report significant results, would achieve the same result (see also, Mayer, 2000).

McCloskey and Ziliak recently updated their earlier study of applied econometric papers published in the American Economic Review during the eighties (McCloskey and Ziliak, 1996) with a similar study for the nineties (Ziliak and McCloskey, 2004). Their results suggest that Denton's "publication filter" might be more than a theoretical possibility even in leading journals: they found that no less than three quarters of the papers published during the nineties used statistical significance as a sole criterion for the inclusion of a variable (Ziliak and McCloskey, 2004: Table 1). Give such a standard it is easy to see why an author might "self-select" not to submit a paper with insignificant coefficients.

Regrettably, the "publication filter" is not the only disincentive for evaluative work that might detect and

discourage unwarranted data mining: there are few career rewards for such work, and, embarking on a replication might be read as a sign of sterile imagination by colleagues. Worse still, an attempted replication might then be viewed as a personal attack. Finally, repetition is hard, especially since the original data set and statistical algorithm is often hard to mimic[14] (Dewald et al., 1986).

In biology, a science where observational data is also regularly used in empirical investigations, it has long been recognised that such data places an extra burden on scientists to avoid misleading results. As the first editor of *Biometrika*, Francis Galton urged the distribution of data sets with papers, where practical, and the creation of a databank where such data may be accessed for critical reworking (Galton, 1901). Ragnar Frisch matched Galton's sentiments when he announced, in the first edition, that *Econometrica* would, normally, publish "raw data" with applied papers to "stimulate criticism, control and further studies" (Frisch, 1933: 3).

While *Econometrica* has not, in the main, found it useful to implement the policy envisaged by Frisch, the *Journal of Money, Credit and Banking* launched the *Data Storage and Evaluation Project* in 1982 to facilitate the evaluation of econometric results[15]. But even there Dewald, Thursby and Anderson (1986) found it remarkably difficult to replicate published results, often due to errors, but also because data and programmes had not in fact been stored.

The optimistic perspective of Pagan and Veall (2000) on the ability of replication and criticism in journals to expose and discourage unwarranted data mining should therefore be tempered with two observations: replication is difficult in economics and the incentives for such research may be weak. As a consequence, replication remains scarce and refutations scarcer still (for example, Greene, 2000).

Aris Spanos once claimed that he knew of no "economic theory that was ever abandoned because it was rejected by some empirical econometric test, nor was a clear-cut decision between competing theories made in lieu of the evidence of such a test" (Spanos, 1986: 660) and Lawrence Summers asked his readers to identify even a single "meaningful hypothesis about economic behaviour that has fallen into disrepute because of a formal statistical test" (Summers, 1991: 130). Because Summer's challenge came with no incentive to respond, Keuzenkamp and Magnus (1995) offered a prize[16] to any reader that was able to offer evidence to disprove Spanos and Summers; nobody won the prize.

Such claims are anecdotal though, and counter-anecdotes exist: Baumol's (1986) case for (an absolute version of) the convergence hypothesis suffered what appears to have been a conclusive refutation at the hands of DeLong (1988). Pagan and Veall (2000) offered further examples to strengthen their claim that the important hypotheses

[14] Dewald et al. (1986) found many authors unable to replicate their own results, notably where data or algorithms had been lost or where a research assistant was no longer at hand to explain her earlier work. In a group of 62 authors in their study, Dewald et al. (1986) found that only 22 provided the requested data and algorithms, while 20 made no reply and the remaining replied that they were unable to comply.

[15] The *Journal of Applied Econometrics* and the *Journal of Business and Economic Statistics* have taken similar initiatives (Pagan and Veall, 2000).

[16] The prize was an all-expenses paid week-long trip to the CentER for Economic Research at Tilburg university.

in economics are examined with vigour and sometimes overturned, such as one important claim by Alesina and the same sceptical Summers[17].

Notwithstanding, there are few strong incentives and many practical hurdles for replication in applied econometrics. It is not surprising that the supply of such papers is modest, making it hard to have confidence that the critical discussion in economics will provide a strong safeguard against unwarranted data mining. Happily recent developments in the theory and practice of econometric modelling might offer additional safeguards.

4. Developments in the theory of econometric modelling

The four recent econometric responses to the challenge of data mining considered in this section move beyond a defence of prevailing practice, to build cases for progressive reform; *progressive*, because these are not calls for a return to stricter rules to avoid data mining. These are examples of the second and third response suggested by Hoover and Perez (2000), according to which data mining is seen as a crucial (or at least an inevitable) part of a sensible modelling strategy that uses the vast computing power offered by modern information technology (Hand, 1998).

All four modern responses start from the shared experience mentioned above: that econometricians grapple with model uncertainty and usually with non-experimental data. However, the direction an econometrician takes from this shared experience depends on how she conceptualises the goal of econometric modelling. Specifically, the different positive approaches to data mining move apart from the shared experience in pursuit of different ontological and epistemological visions of the underlying data generating process (DGP).

Three of these are inductive strategies, the first of which is built on a radical ontology that rejects the very concept of a DGP (section 4.1). The differences between the other three strategies are epistemological, not ontological, with disagreement on the ability of the modelling process to discover the DGP. The DGP moves from the periphery to the centre of the modelling exercise in the final response (section 4.4). In that approach the objective is to uncover the (local) DGP and the beneficent data mining exercise proceeds deductively from a statistical model that the econometrician claims to be a valid reduction of the (local) DGP[18]. This section considers these strategies in turn, and moving along the DGP spectrum (introduced above) from the nihilistic extreme to the opposite extreme and the goal of encompassing the (local) DGP.

4.1 A first inductive response: robust talk

Data mining poses not only a risk to the formal qualities of statistical inference in econometrics, but as a result thereof, also compromises the believability of particular models from the perspective of their consumers: other economists, policy makers and the broader public. Since the econometrician is uncertain about the appropriate

[17] Pagan and Veall and referring to Alesina and Summer's claim that central bank credibility would lower the sacrifice ratio for monetary policy (Pagan and Veall, 2000).

[18] Pagan and Veall have elsewhere asked whether the emphasis on the centrality of the DGP in the G-S approach implies that "other approaches do not have a similar aim" (Pagan and Veall, 2000: 214)? This paper answers Pagan and Veall in the affirmative.

model and has to discover the latter with observational data, the incentive for an applied econometrician who intends to persuade her clients of a model's veracity, is to select one (or a few) preferred estimation(s) iteratively and only to report these. More especially, the many specifications that yield results in conflict with the preferred model go unreported.

These incentives are understood by the customers of econometrics too – or so argued *inter alia* Cooley and LeRoy (1981), Leamer (1983) and, more recently, Mayer (2000) – and these customers are correctly sceptical of applied econometric output; a strategic setting Cooley and LeRoy (1981: 826) described as "…nearly a zero-communication information equilibrium. The researcher has the motive and opportunity to represent his results selectively, and the reader, knowing this, imputes a low or zero signal-to-noise ratio to the reported results". Their claim was followed by Leamer's emphatic assertion that "[h]ardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone's else's data analysis seriously" (Leamer, 1983: 37). At stake is the quality of the scientific discussion in applied economics and not so long ago Peter Phillips saw a "..huge credibility gap that exists between economic theory, empirical evidence and policy prescriptions" (Phillips, 1988: 357).

This risk to the scientific discussion in economics is not simply theoretical, as survey evidence suggest that a large proportion of economists are sceptical of econometric output[19]. The first class of formal responses to the challenge of data mining discussed here is attempt to shift econometrics out of this "zero-communication information equilibrium", i.e. to improve the quality of scientific debate.

The recommendations in this class are so many attempts to remove the potentially deceptive aspects that might undermine applied econometrics. Mayer (2000: 186), for example, sketches an "idealized – and simplified – picture of science" in which different econometricians investigating the same hypothesis with the same data, would emerge, in the manner of the legendary translators of the Septuagint, with the same model. The proliferation of models, and the doubtful power of applied models to conclusively reject empirical hypotheses in economics, show that this idealisation does not obtain in economics. For this reason Mayer (2000) chooses to depict data mining as a communications problem and his proposed remedy is to encourage authors to report more than just the final specification, especially that they report specifications that conflict with the final model, as such information would be valuable to the reader.

Edward Leamer agrees with Mayer that data mining undermines the "atmosphere of econometric discourse" (Leamer, 1983: 43).]. In a hierarchy of statements starting with "truth" at the top, followed by "facts", "opinions" and "conventions" at the bottom, Leamer (1983) argues that we rarely reach as high as "facts" and never in econometric modelling. We bring our opinions to the modelling exercise in econometrics, argued Leamer in this severe interpretation of model uncertainty, and these opinions are "whimsical". Consequently the consumers of econometric output have no confidence in any particular estimated model.

[19] Mayer (1995) cited a survey that found 27% of economists were "quite sceptical" of econometric output published in journals with another 2% comprehensively sceptical and 56% "somewhat sceptical".

Since whimsical results fail to convince either economists or the public, Leamer argued, the focus in econometrics had to shift to producing more "robust" results, i.e. results that do not depend on a narrow range of opinions. And to that end Leamer (1983) recommended that the econometrician calculate the implications of many different possible models and then to report the "extreme estimates" of the estimates for particular variables, the width of which will indicate the fragility of inference based on individual models from the class considered. If we took our model uncertainty more seriously, as Leamer (1983) argued, we would uncover how fragile most of our estimates are, and the humble and transparent communication of these fragile results would clear the air that has been darkened by data mining.

Leamer recognises the shared experience of model uncertainty and non-experimental data identified above. But he combines this with a radical view of the purpose of econometric modelling, in which he rejects any attempt to uncover the underlying DGP. He argues that we do not know the underlying data generating mechanism, we will not discover it *en route* with our modelling, nor is it even useful to assume that such a thing as the "true" DGP exists (Leamer, 1983: 36-38). Woodward (2006) has recently argued that such a "radical subjectivism" about the DGP is necessary for the extreme-bounds approach to offer an attractive modelling strategy[20].

This emphasis on robust correlates with the goal of improving communication follows an inductive argument, but it avoids running into the logical problems associated with induction by the very modesty of the goals: the underlying idea is to update the economist's (possibly whimsical) priors about the issue at hand by learning from the data.

Levine and Renelt (1992) implemented Leamer's strategy on one of the most heavily mined data sets in macroeconomics, the cross section data used in the empirical growth literature. They found, however, that very few of the usual suspects in the literature are "robust" correlates of cross-country growth.

However, Levine and Renelt (1992) made no allowance for the quality of the models that are investigated; all of which were treated equally, as so many opinions. Building on their result, Sala-i-Martin (1997) suggested considering the whole distribution (across rival models) of a parameter by calculating the weighted average of the estimated parameter values and of their variances, across all possible models in which it occurs[21]. He then ran 2 million regressions to cover his conception of all the possible permutations of the possible growth regression. In this way he was able to identify 22 apparently robust variables.

More recently, and working with Doppelhofer and Miller, he used a technique midway between the OLS estimates of the 2 million regressions and a fully Bayesian Model Averaging (Sala-i-Martin et al., 2004). The

[20] Woodward (2006) distinguishes "inferential" robustness (in the Leamer sense) from other potentially relevant concepts of robustness, such as measurement robustness, derivational robustness and causal robustness. Hoover and Perez (2004) showed in practice what Woodward (2006) showed in principle, i.e. that extreme bounds analysis is likely to overstate the "fragility" of econometric results. What is more, a correct model of the DGP is not expected to be inferentially robust against alternative specifications (Pagan and Veall, 2000; Hoover and Perez, 2004).

[21] The weights are proportional to the likelihoods of the separate models, i.e. models with higher likelihood receive a greater weight.

assumptions about the DGP remains as agnostic as before, but they are able to report that: "When we examine the cross-country data usually used by growth empiricists using BACE, we find striking and surprisingly clear conclusions" (Sala-i-Martin et al., 2004: 815). They found 18 significant correlates with economic growth out of a potential 67 and their emphasis remains on clarity of communication, without making any claim about the underlying DGP.

The modesty of their claim is perhaps clearly demonstrated by considering those variables identified as robust correlates of growth in both Sala-i-Martin (1997) and Sala-i-Martin et al. (2004). Of the 18 robust correlates in Sala-i-Martin et al. (2004) 3 had been assumed robust in Sala-i-Martin (1997), i.e. the level of initial income, life expectancy at the start of the period and primary school enrolment at the start of the period. Only 8 other variables were identified as robust by both studies, they were 3 dummy variables (Spanish colony, Latin America dummy, and Sub-Saharan Africa dummy), 3 religion fractions (fraction Confucian, fraction Muslim, fraction Buddhist) and 2 other variables (fraction of gdp in mining and the number of years as an "open" economy). This set is doubtless easy to communicate, but makes no attempt at uncovering anything that might be thought of as "causes", in the way that, for example, Acemoglu and Johnson (2005) have done. But this is not a telling criticism against a radically subjectivist approach that aims only at improving the communication of econometric results in the face of the extensive data mining that is likely to result from model uncertainty and non-experimental data.

4.2. A Second inductive response: The "Reality Check" and RETINA

Halbert White (2000) has been concerned with data mining in econometric models designed for forecasting, and has offered a rigorous test to reduce the risk of confusing skill and luck in such models. Specifically his test is for the null hypothesis that the best model selected by the chosen method does not outperform a given forecasting benchmark model; he calls this test the "Reality Check" (White, 2000).

As stated, the DGP plays little role in White's Reality Check, but the DGP could be introduced along the following lines: White is concerned with identifying the "truly best" forecasting model, and this would match the DGP if the latter was simple enough and time invariant over the relevant data sample and forecasting horizon (not unlike the local DGP in section 4.4 below). However, this is not a step White (2000) felt compelled to take as his focus remained on forecasting performance.

The Reality Check is a systematic approach that protects econometricians from being impressed by spuriously accurate forecasts on mined data, such as the example by Sullivan et al. (2001) discussed in the second section. Such a procedure could be automated and, indeed, White has suggested a relevant algorithm that combines the focus on forecasting with considerations such as flexibility in functional form (allowing for non-linearities and interaction terms), and parsimony. The algorithm is called Relevant Transformation of the Inputs Network Approach (RETINA) (White, 1998) and has since become available for wider use (on GAUSS and MATLAB platforms) (Perez-Amaral et al., 2003).

The four stages of the RETINA algorithm are discussed extensively in Perez-Amaral, Gallo and White (2005) and

Castle (2005). Stage 1 entails preliminaries such as data transformation and a three way splitting of the data into sub-samples. Stage 2 is a step-wise model search using only data from the first sub-sample, but tested in an out-of-sample forecast on the second sub-sample. This leads to a preferred model. Stage 3 uses the second sub-sample to search for a more parsimonious version of the preferred model which will again be tested out-of-sample against the third sub-sample. Finally in stage 4 the algorithm repeats stages 2 and 3, but with the ordering of the sub-samples reversed. The final preferred model will have the best forecast performance over the entire sample.

RETINA is a specific-to-general algorithm which selectively adds variables to a model with the goal of good forecasting ability within a parsimonious model. As such, it is another inductive approach to data mining; but, in contrast with the focus on better communication in the previous method, the emphasis here is on improving out-of-sample forecasts. It is an appeal to the old "gold standard" of econometric modelling, i.e. forecasting, to show that the inevitable data mining yielded a "useful model". And though it would be possible to find a minor role for the DGP in this procedure, that would be step beyond what the proponents of the procedure feel comfortable to take; note, for example, how Perez-Amaral, Gallo and White describe RETINA's goal: "Identify a parsimonious set of (transformed) attributes likely to relevant for predicting out-of-sample" (2005: 266).

4.3 A Third inductive response: automated model selection with PIC

Peter Phillips and Werner Ploberger (for example, Phillips, 1996; Phillips and Ploberger, 1996; Ploberger and Phillips, 2003) have developed a third inductive approach to data based model selection which allowed them also to create an automated data-based modelling algorithm. The DGP takes an even more central role here, as a regulative idea[22], even if the goal of the modelling exercise is not to uncover the DGP.

The statistical foundations for Phillips and Ploberger's approach is the earlier work on stochastic complexity by Rissanen (especially, Rissanen, 1986; 1987). Building on these foundations, Phillips and Ploberger are able to provide a useful (and consistent) modelling strategy as well an epistemological critique of what econometricians can hope to achieve given that "…the true model for any given data is unknown and, in all practical cases unknowable" (Phillips, 2003: C26).

Their agnosticism about the DGP is a tempered version of Rissanen's radical empiricism, about which he leaves little doubt in the reader's mind when he expresses the wish to "… remove the untenable assumption of data generating systems and 'true' parameters, we instead regard the class of models to provide a language in which to express the regular features in the data"[23] (Rissanen, 1986: 1080). Phillips (1996) used this last extract from Rissanen (1986) to introduce his ideas on the DGP, and he has often referred favourably to the concept of a model providing a "language in which to express the regular features of the data" (Phillips, 1996: 766; 2003: C39)
.

[22] Even though this procedure is not intended to uncover the DGP, the merit of any particular model is defined relative to the "unattainable" DGP.

[23] Rissanen expresses the same ideas elsewhere, for example: "In our general philosophy of modelling there are no data generating probabilistic systems nor "true" parameter values" (Rissanen, 1986: 1092).

But Phillips and Ploberger do not follow Rissanen all the way to an ontological rejection of the DGP. In contrast, they use the DGP to derive limiting theorems for the ambitions of econometric modelling that determine "quantitative bounds" or "limits" to how close an actual econometric model can approach the DGP (also called the "proximity bound"). However they attribute the theorem which proves this proximity bond to Rissanen[24]. This limit is a positive function of the number of parameters in the initial model and a function of the nature of the data. Where the latter is concerned Phillips and Ploberger derived the important result that the proximity bound is wider for trending data, i.e. it is harder to find good models for trending data (Ploberger and Phillips, 2003).

The relevance of this theorem for data mining lies in the result that even with infinite data, a model cannot cross over the proximity bound; model uncertainty is not just a small sample problem in this conceptualisation. In contrast, Pagan and Veall (2000) argued that the risks of data mining were largely small sample problems and, more strongly, Hoover and Perez (1999) and authors in the G-S tradition such as Hendry and Krolzig (for example, Hendry and Krolzig, 1999) have used White's theorem to argue that their data mining algorithm would, asymptotically, uncover the DGP with a probability of one.

Phillips and Ploberger's large sample result – that the model can reach but not cross the proximity bound for any given DGP – implies the need for a "yardstick" with which to measure rival models in finite samples. To this end they introduced a Bayesian information criterion, called the Posterior Information Criterion (or PIC) with attractive small and large sample properties; for example, it attains the proximity bound asymptotically (Ploberger and Phillips, 2003).

The posterior information criterion could be used to guide the iterative model selection strategies associated with data mining. Indeed Phillips and Ploberger were also pioneers in the field of automated model selection with the purpose of building good forecasting models, or what they have called "data based automation" (Phillips, 2005: 15). Their data based automation uses the PIC to help the econometrician with difficult, but important decisions, such as the selection of an appropriate lag length, the inclusion or otherwise of an intercept, the specification of a trend and the inclusion and timing of structural breaks (Phillips, 1995b). A very exciting prospect with this algorithm is its potential for offering a wed-based interface which would allow modellers to access the main algorithm via the internet (Phillips, 2005).

In addition to its use as a model selection criterion, the PIC has also been extended for purposes of comparing rival forecasting models, in which case it is called the forecast-encompassing PIC criterion (PICF) (Phillips,

[24] Philips and Ploberger define the "distance" between the model and the underlying DGP in terms of the Kullback-Leibler (KL) distance. The DGP therefore plays a role in their conception of the limit to econometric modelling. Such a conception has no meaning in Rissanen's more radical empiricism, despite Phillips's (2003: c40) claim that Rissanen has "asked how close on average (measured in terms of Kullback-Leibler distance) can we get to a true dgp using observed data". Rissanen's rejection of the KL distance is clear from the following: "The same is true about many other well-know model-selection criteria such as the AIC, where the objective is to estimate either a mean prediction error or the Kullback distance, both of which involve the expectation relative to an imagined and non existing "true" distribution" (Rissanen, 1987: 224).

1995b). Since the PICF is automatically available for each model specified in this way, it is also possible to combine the forecasts of various models by weighting their various forecast by their posterior likelihoods. This method Is not only theoretically attractive, as Phillips has used it with encouraging effect in applications on two well-known data sets in the macroeconometric literature (Phillips, 1995b; a).

Using an information criterion such as PIC as a model selection criterion is precisely how Granger, King and White (1995) suggested the risk of unwarranted data mining might be diminished. They preferred an inductive model selection criterion based on a metric that measures the gap between the model and the DGP, to the deductive model selection criterion based on the theory of reduction which is the topic of the next sub-section and which is built around a set of criteria for an empirical counterpart to the (local) DGP.

4.4. A deductive response: Probabilistic reduction

In the course of twenty-five years the G-S method (also called the LSE method, or sometimes the Hendry method) has risen to take a leading place amongst the rival methods for econometric modelling[25]. It is a tradition often associated with the London School of Economics, especially with David Hendry (though lately at Oxford), and collaborators. The core theoretical contribution of this method is the use of probabilistic reduction theory as a framework for empirical modelling, a theory which conceives of the entire modelling process as a series of reductions from the "unknown high-dimensional distribution" which is called the Data Generating Process (DGP), or alternatively the Haavelmo distribution (1989; 2005b). While probabilistic reduction theory is abstract, the G-S modelling approach is a practical "analogue" or "embodiment" thereof and designed to facilitate data-based econometric modelling (Campos et al., 2005b: 15).

The high dimensionality of the DGP implies that no econometric model could be its empirical counterpart, a view also shared the three alternative approaches to data-based modelling discussed above. The differences between these approaches lie in the next step: while some discard the idea of the DGP, others decide to focus on forecasting and others again try to approach the unattainable DGP.

In contrast with those approaches the G-S literature introduces a new concept to the debate, the local DGP (LDGP): latter represents a valid reduction of the DGP and is a smaller probabilistic model (a well behaved Haavelmo distribution) showing the parameters of the interest for the project at hand. Defining the parameters of interest for a particular project is, in fact, the first of ten steps in the reduction sequence. It is followed by: data transformations and aggregation; sequential factorisation; data partition; marginalization; mapping to stationarity; conditional factorization; constancy; lag truncation and functional form (Campos et al., 2005b). A valid reduction path from the true DGP to a "well behaved" LDGP entails no loss of information.

The econometric model will be a postulated empirical counterpart to this LDGP, and it is connected via the valid

[25] For early surveys see Pagan (1987) and Gilbert (1986), and more recently Campos, Ericsson and Hendry (2005b). Leading texts include Hendry (1995) and Spanos (1986; 1999) and a recent compendium of critical papers in the development of this method has appeared under the editorship of Campos, Ericsson and Hendry (2005a).

steps of reduction with the DGP itself. Once the LDGP has been conceptualised, the implementation of G-S starts with an overparameterised general model (generalised unrestricted model or GUM) which is conjectured to nest the LDGP.

The next step is crucial in the G-S logic, and it is also the critical point where the G-S method dispels concerns with unwarranted data mining (see below). This critical step is to determine whether the GUM provides an statistically adequate approximation to the LDGP, i.e. whether the GUM is a congruent description of the LDGP in Hendry's terms (Hendry, 1995). A dominant (or encompassing) congruent model is one that accounts for the information in (i) the relative past; (ii) the relative present; (iii) the relative future; (iv) information from economic theory, which helps to define the parameters of interest; (v) information about measurements; (vi) information in rival models (again subdivided by relative past, present and future) (Hendry, 1995)[26].

A battery of mis-specification tests are used to establish the congruency of the GUM, and if the GUM is congruent, a series of simplifications are implemented to uncover a parsimonious econometric model without unacceptable loss of information. Each simplification is tested for a loss of information, and whether it leaves the simplified model congruent with the LDGP. These simplifications proceed deductively as a series of restrictions on the GUM which is now treated as congruent with the LDGP. This series of deductive tests and the deductive reduction of the DGP to the LDGP explains why the G-S method is categorised as a deductive in this paper and contrasted with the various inductive approaches considered above.

Critics have raised a number of objections to G-S modelling, and the following paragraphs consider those that are relevant to data mining: first, the layers of testing in this explicitly iterative modelling strategy has exposed the G-S method to the suspicion of unwarranted data mining (Keuzenkamp, 1995; Faust and Whiteman, 1997b; a). Proponents of the G-S method have answered these concerns by, first, distinguishing "warranted" from "unwarranted" data mining (Spanos, 2000) or "constructive" from "pejorative" data mining (Campos and Ericsson, 1999) and, second, arguing that the data mining in G-S is "warranted" or "constructive".

For example, Spanos (2000) argues that a G-S strategy uses the battery of mis-specification tests only once for a postulated GUM after which the GUM is either judged to be congruent, or rejected. Failed mis-specification tests are not treated constructively and are not used to redesign the GUM so as to immunise the GUM against a given test, a step that would have introduced concerns over the respecification bias (discussed above). Instead a rejected GUM is discarded and the econometrician is required to rethink the LDGP before formulating another estimable GUM (Spanos, 2000).

Second, Hoover and Perez (1999) raised the possibility that simplification of a congruent model might be path dependent. In principle the various end-point models could be tested for encompassing, but given the

[26] As mentioned above Granger et al. (1995) preferred a model selection criterion based on a distance measure to the benchmark approach in the G-S approach, because they considered it problematic to select an particular set of qualifying criteria. It might be difficult to convince others that these are reasonable criteria. This line of criticism is countered, in the G-S literature, by building a case for the congruency criteria.

tremendous labour involved in tracing even a single path given a fairly general GUM, this rarely happens in practice.

Third, the G-S method might lead to over-fitting, by including variables which are opportunistically present in the GUM. And, finally, since the simplifications are performed iteratively on the same data set, the test statistics and especially the size of the tests cannot be interpreted in the standard fashion. Hoover and Perez (1999: 169) call the G-S test statistics 'Darwinian', i.e. "the tests statistics for any specification that has survived such a process [of reduction] are necessarily going to be 'significant'. They are 'Darwinian' in the sense that only the fittest survive". There is uncertainty over the size of such "Darwinian" test statistics.

This last criticism has been answered in theory and the last three in practice. First, where theory is concerned the proponents of the G-S method have appealed to a theorem by Halbert White (2005 [1990]) on the asymptotic size and power of a specification-based model selection procedure such as G-S. White (2005 [1990]) shows that a general model that encompasses the LDGP will recover that DGP, with zero type I and II errors as the sample grows to infinity, a result that turns the table on the criticism of Darwinian test statistics. Hoover and Perez (1999) summarised this remarkable result:

> "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" (Hoover and Perez, 1999: 170).

White's theorem provides theoretical encouragement to a strategy based on a (probably) over-parameterised unrestricted model as starting point. But it also seems to contradict the theorem by Rissanen which Ploberger and Phillips used to derive a "proximity bound" between the an over-parameterised starting point and the DGP, a bound which not cannot be crossed even as the sample size grows to infinity.

This contradiction between the theorems of White and Rissanen is striking and it is surprising that it has not attracted much discussion on either side of the literature. The contradiction arises because of the following assumptions by White (and the G-S authors) which are rejected by Rissanen (and Phillips and Ploberger). White and G-S authors assume that an econometric model could, in principal, nest the DGP, while Phillips and Ploberger reject the possibility of such nesting and Rissanen rejects the idea of a DGP. As mentioned, the local DGP concept is critical for the G-S approach and now we see why: it is the reduction from the high dimensional DGP to the relevant LDGP that facilitates both an specification-based strategy such as G-S and the applicability of White's theorem. The theoretical inconsistency between the ambitions of PIC and G-S originates in the theory of reduction.

However, using the theory of reduction and White's theorem to answer the theoretical concerns over Darwinian test statistics leaves unanswered the practical concerns about finite sample properties. To investigate this practical

question Hoover and Perez (1999) translated the G-S method to an algorithm and automated the procedure on a computer. This allowed them to use Lovell's (1983) famous Monte Carlo set-up to test the ability of G-S to discover the known DGP in Lovell's (1983) artificial economy. They could also evaluate the practical relevance of the following three criticisms of G-S: path dependency, over-fitting and unknown test size.

Hoover and Perez's (1999) results were encouraging; in contrast with the model selection criteria studied by Lovell (1983), the automated G-S algorithm recovered the DGP with considerable (though not nearly universal) success and the size and power t-test statistics in the final models were not much distorted by the iterative testing of the G-S procedure. Hendry and Krolzig (1999) responded to Hoover and Perez's (1999) simulation by introducing an automated G-S algorithm of their own, called *PcGets* (Hendry and Krolzig, 2001), with which they were able to improve on Hoover and Perez's (1999) results. The data from Baba, Hendry and Starr (1992) (a widely known earlier paper in the G-S literature) served as real world test for *PcGets*, and again the results were highly encouraging, with a final specification close to Baba et al.'s (1992), but reached within seconds instead of weeks or even months. From Hendry and Krolzig's perspective, a considerable merit of this powerful search algorithm is precisely that it affords the econometrician greater time and resources to "improving the theory, data measurement and econometric specification underpinning the GUM" (Hendry and Krolzig, 2004: 800).

In the G-S tradition, as in the other three traditions discussed above, recent advance in automated modelling have generated powerful tools, that save time and search costs, and that should free more resources for the part of modelling where econometricians add more value. Though the automated G-S algorithms imply large numbers of iterative tests, there is no evidence yet, in theory or practice, that they expose the modeller to larger risks of unwarranted data mining.

5. Data mining races

The relatively easy access to the newly developed automated modelling algorithms have encouraged econometricians to pit them against each other in what may be described as "data mining races". Recent examples include: Hoover and Perez (2004), Hendry and Krolzig (2004), Castle (2005), Perez-Amoral, Gallo and White (2005). There is a sense in which this seems an obvious empirical test of the claims made for the various responses to the risk of unwarranted data mining. The proof of the pudding is in the eating, and with the automated versions of the implied algorithms readily available, the case for empirical trials seems compelling.

A typical example in this literature is Hoover and Perez's (2004) simulation based on Levine and Renelt (1992) cross-country growth data set as the testing ground for two versions of Leamer's extreme-bounds analysis – those of Levine and Renelt (1992) and Sala-i-Martin's (1997) – on the one hand and their own automated version of the G-S method on the other. They also run Sala-i-Martin's (1997) method against their own G-S algorithm on Sala-i-Martin's (1997) original data set.

Hoover and Perez (2004) are well aware that the extreme-bounds approaches entails no concept of the DGP as a

goal of the modelling exercise. Since they don't see any merit in the radical subjectivism of Leamer and others[27] they construct their data mining race as if this method was aimed at uncovering the true DGP. In contrast the G-S method, which is the second competitor in their race, proceeds deductively once the modeller is satisfied on statistical grounds that her general model is congruent with the local DGP.

Hoover and Perez's (2004) results amplify the favourable evaluation of G-S in Hoover and Perez (1999), and this against two versions of extreme bounds analysis: Levine and Renelt's (1992) method is discovered to be too strict, as it eliminates many true variables, while Sala-i-Martin's (1997) is not strict enough, allowing too many opportunistic variables in the final specification. In contrast, the G-S method is "just right" in terms of both test size and power (Hoover and Perez, 2004: 790); a result confirmed by Hendry and Krolzig (2004) using their PcGets algorithm (Hendry and Krolzig, 2001).

While demonstrating the merit of the G-S approach with cross-section data adds important information to Hoover and Perez's earlier demonstrations in the time series context (Hendry and Krolzig, 1999; Hoover and Perez, 1999), it is less clear we can learn from this race and the "future horse races against other search methodologies" which they eagerly anticipate (Hoover and Perez, 2004: 790). The failures of extreme bounds analysis in their tests are not relative to successes that Leamer or other proponents of these methods seek. This same conclusion would have followed a discussion of most papers in this literature. The research question in such a data mining races is not well defined.

6. Conclusions

Tinbergen and the other pioneers of modern econometrics already encountered the central features of our subject matter – model uncertainty and nonexperimental data – which opens the door to unwarranted data mining. And despite the pious intentions of Tinbergen's generation the culture of repetition and criticism of applied econometrics has flourished only modestly. Though the critical discussion in academic journals and competition in the market for econometric output, there is little evidence of vibrant competition et least where the journals are concerned.

But new hope springs from a number of exiting developments in the theory of economic modelling: this paper surveyed four of these developments ranging from a radically subjective inductivism to an objective and structural and deductivism.

The subjectivist approach, associated especially with the work of Edward Leamer, regards data mining as a problem of communications in economic research and the proposed solution (EBA) is designed to expose the fragility of empirical results in econometrics. In contrast, Hall White's "Reality Check" and the RETINA algorithm deal with the risks of data mining by adjusting test for all possible rival models and by using the "gold standard" of econometrics, new data, to test a model.

[27] Elsewhere, Hoover and Perez found Leamer's position on the DGP "barely coherent" (Hoover and Perez, 2000: 201).

The DGP does not feature in either EBA or RETINA, but moves closer to centre stage in the third approach discussed here, Phillips and Ploberger's PIC. This Bayesian criterion is a yardstick to guide a model search, given model uncertainty and non-experimental data. Though the associated modelling strategy does not aim to uncover the DGP, it does aim for (and asymptotically reach) the proximity bound of the DGP. In contrast with the other approaches, G-S is a deductive approach with the explicit aim of encompassing the DGP. The risks of unwarranted data mining are avoided through the application of the theory of reduction: a once-off round of mis-specification tests on the general unrestricted model is followed by a series of deductive tests, the validity of which is based on the statistical adequacy (congruency) of the GUM. There are no invalid inferences, the tests retain their nominal size and consequently there is no risk of unwarranted data mining. G-S, as well as PIC and RETINA can and have been automated and, especially the former, has built an impressive track record of applications.

If Keynes sent 70 modern Tinbergen's into cells with data and laptops they could, following one of the various modelling strategies described here, conceivably emerge with similar models. And they would have done so rapidly using one of the automated algorithms mentioned above. But would they have chosen the same algorithm, and which one is the fittest (if any)?

This question admits of no easy answer: the experience of twenty years and the tremendous gain in computing power have introduced powerful tools that will reduce the cost of the slave-work in applied econometrics, but they have not reduced the importance of the first step in such a project: the research question.

A project's goal should still be the overriding factor in determining the appropriate research strategy (Granger, 1999). Is it forecasting financial variables? Then RETINA looks promising; or perhaps PICF? Is it a structural model for aggregate consumption? Then *PcGets* promises much. And in both cases the initial set-up, including the data set, is critical and there is as yet no algorithm to reduce that task. Indeed Hendry has consistently argued that these data mining algorithms free the econometrician to labour at that end of the modelling effort where he is most able to contribute (for example, Granger and Hendry, 2005).

Amongst the decisions that the econometrician cannot outsource to his data mining algorithm is the philosophical question of his goal relative to the DGP. This issue turns on deeper questions in ontology and epistemology, the problem of induction, and the bridge between theoretical constructs and reality in empirical science. And the data mining races of recent vintage between algorithms built on different configurations of answers to these questions are regrettably uninformative.

References

Acemoglu, D. and S. Johnson (2005). "Unbundling institutions." Journal of Political Economy 115(5): 949-995.

Baba, Y. D., D. F. Hendry and R. M. Starr (1992). "The demand for M1 in the USA, 1960-1988." Review of Economic Studies 59: 25-61.

Backhouse, R. E. and M. S. Morgan (2000). "Introduction: is data mining a methodological problem?" Journal of Economic Methodology 7(2): 171-181.

Baumol, W. J. (1986). "Productivity growth, convergence and welfare." American Economic Review 76(December): 1072-1085.

Burger, R. P. and S. A. du Plessis (2006). Quantifying the extent of data mining in applied econometrics. Stellenbosch, mimeograph.

Campos, J. and N. R. Ericsson (1999). "Constructive data mining: modeling consumers' expenditure in Venequela." Econometrcis Journal 2: 226-240.

Campos, J., N. R. Ericsson and D. F. Hendry, Eds. (2005a). General-to-Specific Modelling. The international library of critical writings in econometrics. Cheltenham, UK, Elgar Reference Collection.

--- (2005b). General-to-specific modelling: an overview and selected bibliography. Washington, Board of Governors of the Federa Reserve System, International Finance Discussion Papers, Number 838.

Castle, J. L. (2005). "Evaluating PcGets and RETINA as automatic selection algorithms." Oxford Bulletin of Economics and Statistics 67(Supplement): 837-880.

Chatfield, C. (1995). "Model uncertainty, data mining and statistical inference." Journal of the Royal Statistical Society, Series A 158(3): 419-466.

Clements, M. P. and D. F. Hendry (2005). "Evaluating a model by forecast performance." Oxford Bulletin of Economics and Statistics 67(Supplement): 931-956.

Cooley, T. and S. LeRoy (1981). "Identification and estimation of money demand." American Economic Review 71(December): 825-844.

DeLong, J. B. (1988). "Productivity growth, comnvergence and welfare: comment." American Economic Review 78(December): 1138-1154.

Denton, F. T. (1985). "Data mining as an industry." The Review of Economics and Statistics 67(1): 124-127.

Dewald, W. G., J. G. Thursby and G. Anderson (1986). "Replication in empirical economics: the Journal of Money, Credit and Banking Project." American Economic Review 76(4): 587-602.

Faust, J. and C. H. Whiteman (1997a). "General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: a translation and critique." Carnegie-Rochester Conference Series on Public Policy 47: 121-161.

--- (1997b). "Rejoinder to Hendry." Carnegie-Rochester Conference Series on Public Policy 47: 191-195.

Feynman, R. P. (1998). The meaning of it all. London, Penguin Books.

Frisch, R. (1933). "Editorial." Econometrica 1(1): 1-4.

Galton, F. (1901). "Editorial: The spirit of Biometrika." Biometrika 1(1): 3-6.

Gilbert, C. L. (1986). "Professor Hendry's Econometric Methodology." Oxford Bulletin of Economics and Statistics 48(3): 283-307.

Giles, J. A. and E. A. Giles (1993). "Pre-test estimation and testing in econometrics: recent developments." Journal of Economic Surveys 7(2): 145-197.

Glymour, C. (2004). "The automation of discovery." Daedalus Winter: 69-77.

Granger, C. W. J. (1999). Empirical modeling in Economics. Cambridge, Cambridge University Press.

Granger, C. W. J. and D. F. Hendry (2005). "A dialogue concerning a new instrument for econometric modeling." Econometric Theory 21: 278-297.

Granger, C. W. J., M. L. King and H. White (1995). "Comments on testing economic theories and the use of model selection criteria." Journal of Econometrics 67: 173-187.

Greene, C. A. (2000). "I am not, nor have I ever been a member of a data-mining discipline." Journal of Economic Methodology 7(2): 217-230.

Hand, D. J. (1998). "Data mining: statistics and more." The American Statistician 52(2): 112-118.

Hendry, D. F. (1980). "Econometrics: Alchemy or science?" Economica 47(188): 387-406.

--- (1995). Dynamic Econometrics. Oxford, Oxford University Press.

--- (1997). "On congruent econometric relations: a comment." Carnegie-Rochester Conference Series on Public Policy 47: 163-190.

--- (2000 [1985]). Monetary economic myth and econometric reality. Econometrics: Alchemy or science? Essays in econometrics methodology. D. F. Hendry. Oxford, Oxford University Press.

Hendry, D. F. and H. M. Krolzig (1999). "Improving on 'Data mining reconsidered'." Econometrics Journal 2: 202-219.

--- (2001). Automatic econometric model selection using PcGets 1.0. Harrow, Allstar Services.

--- (2004). "We ran one regression." Oxford Bulletin of Economics and Statistics 66(5): 799-810.

Hoover, K. D. and S. J. Perez (1999). "Data mining reconsidered: encompassing and the general to specific approach to specification search." Econometrics Journal 2: 166-191.

--- (2000). "Three attitudes to data mining." Journal of Economic Methodology 7(2): 195-210.

--- (2004). "Truth and robustness in cross-country growth regressions." Oxford Bulletin of Economics and Statistics 66(5): 765-798.

Karni, E. and B. K. Shapiro (1980). "Tales of horror from ivory towers." Journal of Political Economy 88(1): 210-212.

Kennedy, P. (2003). A guide to econometrics (5th edition). Oxford, Blackwell publishing.

Keuzenkamp, H. A. (1995). "The econometrics of the Holy Grail - A review of Econometrics: alchemy or Science? Essays in Econometric Methodology." Journal of Economic Surveys 9(2): 233-248.

Keuzenkamp, H. A. and J. R. Magnus (1995). "On tests and signifcance in econometrics." Journal of Econometrics 67(5-24).

Keynes, J. M. (1940). "On a method of statistical business-cycle research. A comment." Economic Journal 50(197): 154-156.

Krämer, W., H. Sonnberger, J. Maurer and P. Havlik (1985). "Diagnostic checking in practice." Review of Economics and Statistics 67(1): 118-123.

Leamer, E. E. (1978). Specification searches: ad hoc inference with nonexperimental data. New York, John Wiley.

--- (1983). "Let's take the con out of econometrics." American Economic Review 73(1): 31-43.

Levine, R. and D. Renelt (1992). "A sensitivity analysis of cross-country growth regressions." American Economic Review 82(4): 942-963.

Lovell, M. C. (1983). "Data mining." The Review of Economics and Statistics 65: 1-12.

Mayer, T. (1995). Doing economics. Aldershot, Edward Elgar.

--- (2000). "Data mining: a reconsideration." Journal of Economic Methodology 7(2): 183-194.

Mayo, D. G. (1996). Error and the growth of experimental knowledge. Chicago, University of Chicago Press.

McCloskey, D. N. and S. T. Ziliak (1996). "The standard error of regressions." Journal of Economic Literature 34: 97-114.

O'Hagen, A. (1995). "Discussion of the paper by Chatfield." Journal of the Royal Statistical Society, Series A 158(3): 460.

Pagan, A. R. (1987). "Three econometric methodologies: a critical appraisal." Journal of Economic Surveys 1(1): 3-24.

--- (2003). Reflections on some aspects of macro-econometric modelling. Stellenbosch, Keynote address delivered at the 8th annual AES Conference, July 2003, Stellenbosch, South Africa.

Pagan, A. R. and M. R. Veall (2000). "Data mining and the econometetrics industry: comments on the papers of Mayer and Hoover and Perez." Journal of Economic Methodology 7(2): 211-216.

Perez-Amaral, T., G. M. Gallo and D. White (2005). "A comparison of complementary automatic modeling methods: RETINA and PcGets." Econometric Theory 21: 262-277.

Perez-Amaral, T., G. M. Gallo and H. White (2003). "A flexible tool for model building: the Relevant Transformation of the Inputs Network Approach (RETINA)." Oxford Bulletin of Economics and Statistics 65: 821-838.

Phillips, P. C. B. (1988). "Reflections on econometric methodology." The Economic Record(December): 344-

359.

--- (1995a). "Automated forecasts of Asia-Pacific economic activity." Asia Pacific Economic Review 1: 92-102.

--- (1995b). "Bayesian model selection and prediction with empirical applications." Journal of Econometrics 69: 289-331.

--- (1996). "Econometric model determination." Econometrica 64(4): 763-812.

--- (2003). "Laws and limits of econometrics." Economic Journal 113(March): C26-C52.

--- (2005). "Automated discovery in econometrics." Econometric theory 21: 3-20.

Phillips, P. C. B. and W. Ploberger (1996). "An asymptotic theory of Bayesian inference for time series." Econometrica 64: 581-413.

Ploberger, W. and P. C. B. Phillips (2003). "Empirical limits for time series econometric models." Econometrica 71(2): 627-673.

Popper, K. R. (1992 [1961]). The logic of the social sciences. In search of a better world: lectures and essays from thirty years. K. R. Popper. London, Routledge.

--- (2000 [1959]). The logic of scientific discovery. London, Routledge.

Rissanen, J. (1986). "Stochastic complexity and modeling." The Annals of Statistics 14(3): 1080-1100.

--- (1987). "Stochastic complexity." Journal of the Royal Statistical Society, Series B 49(3): 223-239.

Sala-i-Martin, X. (1997). "I just ran two-million regressions." American Economic Review (Papers and Proceedings) 87(2): 178-183.

Sala-i-Martin, X., G. Doppelhofer and R. I. Miller (2004). "Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach." American Economic Review 94(4): 813-835.

Schumpeter, J. A. (1933). "The common sense of econometrics." Econmetrica 1(1): 5-12.

Spanos, A. (1986). Statistical foundations of econometric modelling. Cambridge, Cambridge University Press.

--- (1989). "On rereading Haavelmo: a restrospective view of econometric modelling." Econometric Theory 5:

405-429.

--- (1995). "On theory testing in econometrics. Modeling with nonexperimental data." Journal of Econometrics 67: 189-226.

--- (1999). Probability theory and statistical inference. Cambridge, Cambridge University Press.

--- (2000). "Revisiting data mining: 'hunting' with or without a license." Journal of Economic Methodology 7(2): 231-264.

Sugden, R. A. (1995). "Discussion of the paper by Chatfield." Journal of the Royal Statistical Society, Series A 158(3): 461-464.

Sullivan, R., A. Timmerman and H. White (2001). "Dangers of data mining: the case of calender effects in stock returns." Journal of Econometrics 105: 249-286.

Summers, L. (1991). "The scientific illusion in empirical macroeconomics." Scandinavian Journal of Economics 93(2): 129-148.

Wallace, T. D. and V. G. Ashar (1972). "Sequential methods in model construction." The Review of Economics and Statistics 54(2): 172-178.

White, H. (1998). Artificial neural network and alternative methods for assessing naval readiness. San Diego, NRDA technical report.

--- (2000). "A reality check for data snooping." Econometrica 68(5): 1097-1126.

--- (2005 [1990]). A consistent model selection procedurebased on m-testing. General-to-Specific Modelling volume I. J. Campos, N. R. Ericsson and D. F. Hendry. Cheltenham, Elgar Reference Collection.

Woodward, J. (2006). "Some varieties of robustness." Journal of Economic Methodology 13(2): 219-240.

Ziliak, S. T. and D. N. McCloskey (2004). "Size matters: the standard error of regressions in the American Economic Review." The Journal of Socio-Economics 33: 527-546.