2005

Levels, differences and ECMs - Principles for improved econometric forecasting

PG Allen

R Fildes

Recommended Citation

Allen, PG and Fildes, R, "Levels, differences and ECMs - Principles for improved econometric forecasting" (2005). *Oxford Bulletin of Economics and Statistics*. 7.
10.1111/j.1468-0084.2005.00144.x

Levels, Differences and ECMs – Principles for Improved Econometric Forecasting*

P. Geoffrey Allen† and Robert Fildes‡

†*Department of Resource Economics, University of Massachusetts, Amherst, MA USA (e-mail: allen@resecon.umass.edu)*
‡*Department of Management Science, Lancaster University, Lancaster, UK (e-mail: r.fildes@lancaster.ac.uk)*

Abstract

Unit-root testing can be a preliminary step in model development, an intermediate step, or an end in itself. Some researchers have questioned the value of any unit-root and cointegration testing, arguing that restrictions based on theory are at least as effective. Such confusion is unsatisfactory. Needed is a set of principles that limit and define the role of the tacit knowledge of the model builders. In a forecasting context, we enumerate the various possible model selection strategies and, based on simulation and empirical evidence, recommend using these tests to improve the specification of an initial general vector autoregression model.

I. Introduction

What role, if any, should unit-root and cointegration testing have in a model-development strategy designed for forecasting? Ideally, for a practitioner, principles would be available, amounting to cook-book instructions, on how such tests can best be used in model building. Dharmapala and McAleer (1996) define methodology, when applied to model building, as the 'philosophical basis for the validation and justification of econometric procedures'. Pagan (1987) more explicitly argues that a methodology 'should provide a set of principles to guide work in all its facets', where he interprets 'methodology' to mean a

*We thank the editors and two anonymous reviewers, who convinced us that we had not made our arguments with sufficient clarity.

JEL Classfication numbers: C22, C52, C53.

coherent collection of inter-related methods together with a philosophical basis for their justification and validation. He later complains that econometric model building is overly reliant, not just on the methodology adopted by the modellers, but also on the tacit understanding of its implications as well as personal knowledge and skills (Pagan, 1999, p. 374). If, within a particular methodological approach, principles were available, then such instructions would limit the requirement for the expert's tacit (and personal) knowledge.

It proves to be quite challenging to state and defend a set of clear and operational principles for econometric modelling (Magnus and Morgan, 1999a; Allen and Fildes, 2001; Kennedy, 2002 and the discussion therein), a reflection of the considerable ambiguity in the established literature, and there is certainly nothing that attains the completeness of a cook book, even within a particular model-building methodology. We examine here only a limited subset of issues: those concerned with the utility of the fast-expanding literature on unit-root testing and cointegration analysis when the context is one of improving forecast accuracy. Other aspects of model specification are important and have been considered elsewhere. Choice of initial specification of a general model, data transforms and initial lag order determine the final model. Restricting the initial lag order based on appropriate tests is widely agreed to improve forecast accuracy (Allen and Fildes, 2001).

Not all econometric methodologies embrace unit-root and cointegration analysis with equal facility or enthusiasm. (See, e.g. Leamer's, 1999, p. 150, dismissive remarks.) In fact, Darnell and Evans (1990) treat cointegration analysis as a separate methodology. However, the general-to-specific modelling approach that Pagan (1987) refers to as the LSE Methodology or LSEM (after the London School of Economics where much of the early thinking took place) naturally includes these concepts as potentially contributing to a final model specification.

The aim of this paper is to establish a set of operational principles helpful in model specification based on unit-root and cointegration tests. The modelling framework we adopt is the LSEM. We develop principles by comparing the recommendations from the literature as to how the results of the tests point to alternative model-simplification strategies. They will be based on the comparative empirical and simulation evidence on forecasting accuracy when alternative models are specified in levels, as error-correction models (ECMs) or in differences.

The structure of the paper is as follows. In section II we argue for the need for explicit rules of modelling that would seldom eliminate the need for modeller expertise but instead establish a core of agreed upon, empirically effective principles beyond which expert modellers could contribute. Section III describes potential strategies for building vector autoregressive (VAR) models within the LSEM framework, posing the question as to which of the alternatives tend to produce the most accurate forecasts and under what circumstances. References to the literature on econometric forecasting provide no clear guidance on the choice of modelling strategy as the evidence presented in section IV shows. Nevertheless various simulation studies point to those situations where accuracy improvements may be found. Empirical comparative forecasting-accuracy studies that report the performance of two or more specifications are then shown to give qualified support to those strategies that test for unit roots and cointegration (section V). Structural breaks complicate the picture and represent an active area of research in that the forecaster's identification of such breaks, in the recent past and over the forecast horizon, conditions the strategy to be adopted. The paper concludes (section VI) by stating clear operational principles that have both theoretical and empirical support in leading to improved forecasting accuracy. But Pagan's (1999) complaint still holds – the evidence we found is overly limited and sometimes contradictory, which emphasizes the need for research centred around establishing operational principles of econometric model building and delineating the more limited role of tacit knowledge.

II. The need for principles in econometric forecasting

There are substantial disagreements between econometricians as to how an appropriate model (with a specific purpose such as forecasting in mind) should be developed. This would not matter if methodologies were well defined, and therefore transmissible to others, and gave similar results. Such is not the case. Different groups of econometricians, given a defined data set, following different methodologies, are unlikely to come up with the same model. As evidence, Magnus and Morgan (1999a) persuaded five groups of econometric researchers to forecast the demand for food and obtained substantially different results. Even within the same methodology, the models and corresponding forecast results can differ substantially. Various reasons can be suggested, including the theoretical framework selected and data pre-processing, variability derived from the software and the competence (or otherwise) with which it is employed. But a critical component, even when research groups work within the same broad methodological framework, is the extent to which tacit and personal knowledge affects the operational deployment of the methods subsumed in the methodology (Magnus and Morgan, 1999c, p. 302).

The limited consensus as to how to specify an econometric forecasting model was underlined in a second experiment organized by Magnus and Morgan (1999a) where a novice researcher attempted to develop three different models using the principles embodied in Pagan's three methodologies. This again demonstrated a heavy reliance on tacit knowledge (as well as personal knowledge and skills) and a limited ability to follow the guidance given by the writings of the 'masters' in the particular methodologies. The conclusion we draw from the two prongs of Magnus and Morgan's research is that econometric forecasts and, by implication, their comparative accuracy, are heavily influenced by the choice of methodology made by the research group, the explicit principles that define the methodology's canon, the group's expertise (by which we mean the transmissible and explicit knowledge base used) and their personal knowledge (which cannot be communicated).

Outcome feedback, whereby the results of different model-building processes (and the modellers behind them) can be measured and compared, has the potential of reducing researchers' reliance on the tacit knowledge embodied in the application of a methodology. It has been little used in econometrics as it applies to forecasting. Instead researchers have used self-referential, often asymptotic, statistical arguments as the sole justification for the procedures adopted. In contrast, time-series statisticians have employed so-called forecasting competitions (Fildes and Ord, 2002) to evaluate both the methods and the tacit knowledge of the statistician forecaster. For example, in the M-2 Competition, and in the comparisons of personalized autoregressive integrated moving average (ARIMA) identification procedures vs. automatic procedures, the value added by the forecaster's personal knowledge has been appraised. The role of a principle where outcome feedback is available is therefore to define the added value that expertise brings to these different modelling approaches.

Undoubtedly, specifying principles in multivariate analysis will be more difficult than in univariate settings. The reason, Magnus and Morgan (1996b, p. 376) argue is that, in model specification, universal principles are hard to establish where a complex 'combination of circumstances are involved so that no simple, single-circumstance, textbook rule' can be invoked. Nevertheless, stating straightforward conditional principles is possible and these could still be enhanced by the tacit knowledge of the researcher.

Within the LSEM framework, testing for unit roots and cointegration is seen as making a major contribution to model specification. Unfortunately, as Pagan (1999) makes clear, in the context of the LSEM, the 'art-to-science ratio is at an uncomfortable level' and this makes it hard to learn from the writings of master practitioners who may of course disagree on the principles defining the methodology among themselves. This is further confused as the methodology develops over time. Thus, the practice of model building for the purposes of forecasting would benefit from an explicit set of principles that embody the accepted core of the methodology. In ideal form, these principles can be embedded in a model-selection computer algorithm in much the same way as personalized identification of ARIMA models has been replaced by programmed identification routines (Hoover and Perez, 1999; Hendry and Krolzig, 2003a). The development of such programmes allows us to benchmark master practice, identifying just where differences of operational practices appear and therefore the effects (positive or negative) of personal knowledge. The question is how far ‘tacit knowledge can be turned into [principles] and how such rules can be integrated into practice’ (Magnus and Morgan, 1999b, p. 375). Our hope (shared with Pagan, 1999, p. 374) is that, the contribution of communicable explicit knowledge to the results of applied work is high.

In our search for principles on how to use unit-root and cointegration tests, the initial search started with an examination of the econometric texts, which as we note, shows substantial disagreement as to the role of cointegration and unit-root testing in model specification. To reconcile the disagreements, we have then examined the empirical evidence for consistencies. We gave greater credibility to some types of evidence over others. As our concern is forecasting accuracy (measured out of sample), empirical evidence that examines the comparative performance of alternative approaches to achieving a final model specification are accorded the greatest weight. Simulation evidence is also valued but of course usually begs the core question of the relationship of the simulated world to the experienced world. Theoretical and asymptotic arguments are discounted; while they are invaluable as signposts towards establishing a tentative principle, they do not provide any evidence as to operational effectiveness.

III. Model-building strategies

Within a broadly defined LSEM, the specification search starts with a general model compatible with any theoretical model (of the system of interest) deemed appropriate. In practice, given the usual data limitations, the starting point is based on a slight amplification of a model acceptable to the researcher, frequently drawn from the recent literature. This initial model contains additional variables deemed relevant according to economic theory. It contains lags of the variables based on the researcher’s judgment, for which economic theory is usually no guide. Even then, this ‘local data generating process’ can only approximate the complexities of the real economic system; theoretically important variables may be unobservable, unique events may temporarily dominate the stable economic processes being examined, etc. As Phillips (2003) forcefully argues, the ‘true model’ or data generating process (DGP) is both unknown and unknowable. However, a good local model should show congruence within the sample data. Congruence requires that the model match the data in all measurable respects (homoscedastic disturbances, weakly exogenous conditioning variables, constant parameters, etc.; Hendry, 1995; Clements and Hendry, 1998, p. 162).

Within the LSEM, the art of model specification is 'to seek out models that are valid parsimonious restrictions of the general model and that are not redundant in the sense of having even more parsimonious models nested within them that are also valid restrictions of the completely general model' (Hoover and Perez, 1999). This approach to model specification is by no means universally accepted, see, e.g. Kennedy (2002) or Keuzenkamp and McAleer (1995; p. 16), who state: 'Testing downwards is sensible if one favors parsimony, but the theory of reduction does not offer satisfactory principles of simplicity.' However, here we focus on the LSEM with a view to understanding the empirical consequences of alternative reduction strategies within that particular methodology.

There are several strategies for building multivariate equations or systems of equations. Some strategies use unit-root and cointegration tests at various points, and others do not. One possible taxonomy of the strategies is summarized in Figure 1. Unfortunately, where the use of unit roots and cointegration is concerned 'Experts differ in the advice offered for applied work' (Hamilton, 1994, p. 652). In fact, experts, at least those who write books on the subject, seem unwilling to offer much explicit advice at all. Hendry is an obvious exception (Hendry, 1995, 2002).

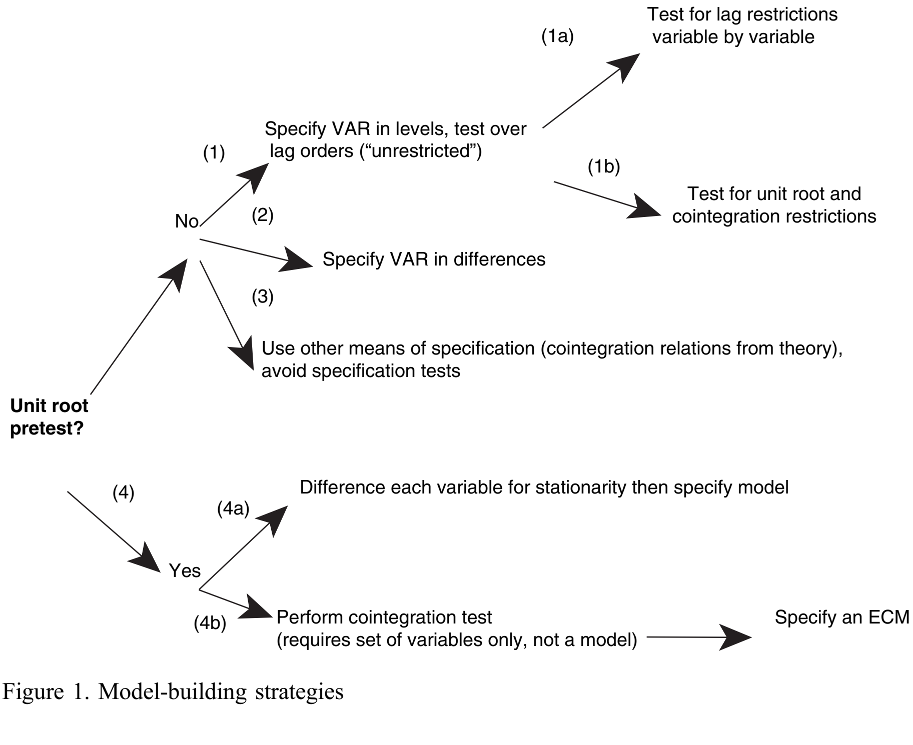
Unit root pretest?
No
Yes
(1)
Specify VAR in levels, test over lag orders ("unrestricted")
(1a)
Test for lag restrictions variable by variable
(1b)
Test for unit root and cointegration restrictions
(2)
Specify VAR in differences
(3)
Use other means of specification (cointegration relations from theory), avoid specification tests
(4)
(4a)
Difference each variable for stationarity then specify model
(4b)
Perform cointegration test (requires set of variables only, not a model)
Specify an ECM

Figure 1. Model-building strategies

A general-to-specific approach usually starts with either a VAR system in levels or a single autoregressive distributed-lag equation. Having started with a consciously over-general model, simplification will need to rely heavily on parameter restrictions, derived from both theory and the data. In addition, the model builder must specify the functional form. Data transformations such as forming ratios, powers, logarithms, or differences can all be thought of as imposing parameter restrictions in models nonlinear in parameters. Adequacy of the initial formulation may be assessed by mis-specification tests, but this does not guarantee that the good causal model is nested within it if the initial model is not sufficiently general. As Hendry (2002) has argued forcibly, theory alone is an incomplete basis for achieving an operational data-congruent model – such an approach, starting with a simple theory-based model, usually has *ad hoc* statistical fixes forced upon it. For the general-to-specific modelling strategy to be successful in balancing over-parameterization with mis-specification, what is required is a reduction strategy that will lead to a good forecasting model.

Strategy 1, referred to as 'specify "unrestricted" VAR in levels' in Figure 1, is simply to reduce the lag order on all variables by 1 and test if the restriction is binding (by a likelihood-ratio test). Repeat until the restriction is binding. For a system with n variables, each reduction in lag order reduces the number of parameters to be estimated by n^2. Test that the final VAR is well specified (based on tests on residuals). Empirical evidence supports this practice (Allen and Fildes, 2001). This is a sequence of pretests each usually conducted at the standard 5% significance level.

A continuation of the strategy, that usually represents a termination point, is strategy 1a: a 'restricted' VAR in levels. This calls for reducing the lag order on individual variables in an unrestricted VAR (e.g. by Hsiao's method, brute force search using Akaike information selection criterion, general-to-specific modelling using PcGets$^{\text{TM}}$; Hendry and Krolzig, 2003b; Owen, 2003). Test that the final restricted VAR is well specified (based on tests on residuals).

Not incompatible with strategy 1a, although usually performed instead of it is strategy 1b: 'post-testing' for unit root and cointegration restrictions. Impose parameter restrictions by performing unit-root and cointegration tests to determine the number and specification of cointegrating vectors to add to each equation in the system. A VAR in levels can be rearranged and reparameterized into a generalized error-correction form with the same number of parameters (an ECM with as many cointegrating vectors as variables).

If parameters on all the error-correction terms in the generalized error-correction form are set equal to zero this corresponds to another strategy: 'Estimate a VAR in differences' (or DVAR, strategy 2 in Figure 1). In practice, difference all variables, then follow the procedure in strategies 1 and 1a. Experts have been unwilling to recommend strategy 2, although Hamilton

(1994) suggested it as one among several possibilities, and Siegert (1999) (in an attempt to apply the LSEM in modelling the demand for food) adopted this automatically, much to Hendry's disgust (Hendry, 1999). But Hendry (1997) himself has noted that when there are structural breaks, a model that is robust to breaks will tend to produce better forecasts. Differencing variables imparts robustness, implying that there are conditions when strategy 2 will be the best.

Cointegration requires variables with unit roots. It also implies parameter restrictions and these are usually across equations. Strategy 3, 'no test,' relies on theory to suggest that variables should be cointegrated, and imposes that specification initially. Similarly, if there are theoretical or historical grounds for expecting a variable to be stationary, such as unemployment rate, there is no reason to difference it and no reason to test for stationarity.

Harvey is one of the strongest proponents of strategy 3. He observes (Harvey, 1997, p. 196): '[M]uch of the time, it [unit-root testing] is either unnecessary or misleading, or both'. As well as doubting the value of unit-root testing, Harvey has little enthusiasm for either vector autoregressions or their modification to embody cointegration restrictions, in part because the modelling strategies (1a, 1b) depend on tests with poor statistical properties. He continues (p. 199):

> However, casting these technical considerations aside, what have economists learnt from fitting such models? The answer is very little. I cannot think of one article which has come up with a co-integrating relationship which we did not know already from economic theory. Furthermore, when there are two or more co-integrating relationships, they can only be identified by drawing on economic knowledge. All of this could be forgiven if the VECM provided a sensible vehicle for modeling the short run, but it doesn't because vector autoregressions confound long run and short run effects.

Diebold (1998, p. 260) makes much the same point.

Probably the commonest strategy is strategy 4, 'unit-root pretest,' where the first step in the analysis is to learn something about the variables of interest by performing unit-root tests on the original variables. A testing strategy is required to determine whether drift (intercept) or deterministic trend or both or neither is present in the series, as the power of the test is reduced by including these terms when the process is not actually present and by omitting the terms when they are needed. The strategy is quite complex although a simpler procedure is available, utilizing prior knowledge about the series (Elder and Kennedy, 2001).

With the information gained in following strategy 4, the researcher can proceed to strategy 4a, 'model with stationary variables'. Difference the non-stationary variables and estimate a mixed VAR with all variables transformed to stationarity. A large number of experts appear to suggest this approach, even though it is not widely practised (and never has been, as far as we can tell).

What we might call ‘early Harvey’ (Harvey, 1990, p. 390) appears to favour this strategy:

> Before starting to build a model with explanatory variables, it is advisable to fit a univariate model to the dependent variable. . . . it provides a description of the salient features of the series, the ‘stylized facts’ . . . An initial analysis of the potential explanatory variables may also prove helpful. . . . In particular the order of integration of the variables will be known. It is not difficult to see that, if the model is correctly specified, the order of integration of the dependent variable cannot be less than the order of integration of any explanatory variable. This implies that certain explanatory variables may need to be differenced prior to their inclusion in the model. A further point is that if the order of integration of the dependent variable is greater than that of each of the explanatory variables, a stochastic trend component must be present.

Hamilton (1994, p. 652) also recommends it, as does Diebold (1998, p. 254): ‘In light of the special properties of series with unit roots, it is sometimes desirable to test for their presence, with an eye towards the desirability of imposing them, by differencing the data, if they seem to be present’. More recently, Stock and Watson (2003, pp. 466–467) conclude:

> The most reliable way to handle a trend in a series is to transform the series so that it does not have a trend. . . . Even though failure to reject the null hypothesis of a unit root does not mean the series has a unit root, it still can be reasonable to approximate the true autoregressive root as equaling one and therefore to use differences of the series rather than its levels.

Probably the commonest strategy of all is strategy 4b, ‘unit root and cointegration pretest’ where the variables found to be $I(1)$ [and in some studies, first differences of variables found to be $I(2)$] are subjected to cointegration testing. Impose the parameter restrictions (if any) that follow from the testing, and estimate the resulting ECM. If no cointegrating vectors are detected, estimate a DVAR. This strategy is followed by ModelBuilder (Kurcewicz, 2002). Holden appears to favour strategy 4b, if cointegration is found to exist, otherwise strategy 4a (Holden, 1995, p. 164): ‘When the variables are not stationary . . . [and if] they are not cointegrated the correct approach is to transform the variables to become stationary . . . and then estimate the VAR in the usual way.’

Considering the popularity of strategy 4b, the outright disregard of it or lack of enthusiasm for it displayed by experts is rather surprising. For example, Maddala and Kim (1998, p. 146) state:

> [I]t is important to ask the question (rarely asked): why are we interested in testing for unit roots? Much of this chapter (as is customary) is devoted to the question ‘How do we use unit root tests?’ rather than ‘Why unit root tests?’ . . . One answer is that you need the unit root tests as a prelude to cointegration analysis. . .

Figure 1 shows a number of strategies for reaching a model specification. Experts offer support for almost all of them. Readers familiar with the lack of consensus among econometricians will be unsurprised. The biggest surprise is that the most widely used strategy (4b) receives so little support.

IV. Theoretical justification and simulation evidence

In theory, when a restriction is true, it should be imposed, as one source of estimation error is removed. One question, and the source of a vast literature, is the ability of a unit-root test or cointegration test to reliably answer whether or not a proposed restriction is true.

Unit-root tests

Several Monte Carlo studies have compared the size and power of various unit-root tests. We are aware of only one that specifically addressed the question we pose here: whether unit-root tests are useful diagnostic tools for selecting forecasting models (Diebold and Kilian, 2000). Diebold and Kilian (2000) conclude that with a data generating process (DGP) that contains roots close to unity, and ‘close to’ probably means ≥ 0.97, a unit-root pretest will signal the presence of a unit root and imposing a unit root will improve forecast accuracy. There are a number of caveats to this finding. The authors used the unit-root test as originally proposed by Dickey and Fuller (1979). They assumed the simplest possible process that contains both deterministic and stochastic trends:

$$y_t - 7.3707 - 0.0065t = \rho(y_{t-1} - 7.3707 - 0.0065(t-1)) + \varepsilon_t,$$

where $\varepsilon_t \sim N(0,\, 0.0099^2)$. The process was intended to mimic US quarterly real gross national product. When $\rho = 1$ (unit root) this gives the random-walk plus drift model, for $\rho = 0$ gives the (deterministic) linear trend model, and for values between gives a mixture of the two models.

As Diebold and Kilian (2000) note, before practical recommendations can be drawn from their study, we need to know whether the results hold for other test procedures, for more complex processes, when lag order is unknown, in multivariate settings, with structural breaks. It is also unlikely that the 5% level of significance is optimal under all situations for deciding when to impose a unit root.

Systems of equations with unit roots

Table 1 compares the theoretical arguments and simulation evidence regarding the imposition of unit roots and the effect of forecast accuracy.

TABLE 1

Out-of-sample forecast accuracy results: Monte Carlo studies and theoretical expectations compared

	Forecast error increases with horizon	
	Without limit	*To a ceiling*
ECM vs. VAR is:		
Always better		
Improving	TH coint, TH nonsta, LT4, LT5, RA, CH	
Worsening		
No pattern		
Always worse		
Improving		
Worsening		TH stat, LT1, LT2
No pattern		
No pattern or varies		
Improving	EY	
Worsening		
No pattern	LT3	
DVAR vs. VAR is:		
Always better		
Improving	TH nonsta, LT3, LT5	
Worsening		
No pattern		
Always worse		
Improving		
Worsening		TH stat, LT1, LT2
No pattern		
No pattern or varies		
Improving	TH coint LT4, RA, CH	
Worsening		
No pattern		
DVAR vs. ECM is:		
Always better		
Improving	TH nonsta, LT3 LT5	
Worsening		
No pattern	LL using HEGY test	
Always worse		
Improving	LT4, RA, CH	
Worsening	TH coint	TH stat, LT1, LT2
No pattern	LL against true	
No pattern or varies		
Improving		
Worsening		
No pattern		

Notes: VAR vector autoregression in levels, ECM unit-root restrictions imposed, DVAR unit-root restrictions imposed on all variables (first differenced).

Forecast accuracy usually measured as TMSFE, the trace of the mean squared forecast error. 'Always better' means that the forecast error from the first model is less than the forecast error from the second model for all horizons, opposite for 'always worse' and 'no pattern or varies' indicates either a single or multiple switches between better and worse. 'Improving' means that the forecast error from the first model consistently becomes smaller relative to the accuracy or the second model as the forecast horizon increases, 'worsening' means the opposite and 'no pattern' means no consistent direction of change.

TH is the theoretical expectation for each comparison, and it depends on whether the DGP is stationary, cointegrated, or non-stationary. With a stationary series, the forecast errors can be expected to increase with horizon towards some ceiling. With a non-stationary series, the forecast errors can be expected to increase without limit.

The Monte Carlo findings are indicated as follows: LTi is the *i*th model of Lin and Tsay (1996); models 1 to 3 are stationary with successively higher characteristic roots, model 3 would likely be identified as containing two unit roots in a standard unit-root test, model 4 contains two unit roots and two cointegrating vectors, model 5 contains four unit roots (and so is difference stationary); RA is the model of Reinsel and Ahn (1992) and except for parameterization is identical to Lin and Tsay's model 4; EY is the model of Engle and Yoo (1987), which is similarly identical to CH, the model of Clements and Hendry (1995), except that Engle and Yoo estimate VAR in levels and error-correction models only, while Clements and Hendry estimate all three; LL summarizes the overall results of Lyhagen and Löf (2003) over seven different DGP, who only report averages of 1–12 steps ahead so that pattern cannot be determined, 'against true' is accuracy of DVAR compared with forecasts from the true DGP and 'using HEGY test' compares DVAR forecasts against the model with cointegrating vectors detected by the HEGY test.

The pattern of forecast errors is divided into those that increase without limit as the forecast horizon increases, the behaviour expected for non-stationary series, and those that increase to a ceiling, the behaviour expected for stationary series (Lin and Tsay, 1996). The category 'always better' indicates that over the entire forecast horizon reported by the study, the first model listed was always more accurate than the second model listed; for 'always worse' the converse is true. The 'no pattern or varies' category usually describes a series, the forecast errors of which switch from favouring one model to favouring the other as the horizon increases, and the subclass 'improving' or 'worsening' shows the direction of change. The 'no pattern' subclass usually indicates a series where forecast errors are similar, regardless of the model.

We first summarize the findings shown in Table 1, before turning to detailed comparisons of the simulation studies. Lin and Tsay's (1996) study is the most comprehensive non-seasonal analysis. Their simulations for clearly stationary series (models 1 and 2 denoted by LT1 and LT2 respectively) conform with theoretical expectations. Imposing any restrictions is a misspecification and more restrictions make accuracy worse.

Where there are groups of variables that cointegrate, restrictions that specify one or more cointegrating vectors should give a better result than either a VAR in differences or a VAR in levels. Imposing restrictions that are not true, for example, estimating an ECM with one cointegrating vector when there should be two or three, should give a worse result than estimating the more general model. Lin and Tsay's (1996) model with two cointegrating vectors (LT4) supports this. Interestingly, all studies that assume a cointegrated DGP find that estimation of a DVAR gives improving – though not better – forecasts at longer horizons.

Failing to impose restrictions, and estimating a VAR in levels when the DGP is cointegrated, should be less efficient, but harmless and so the VAR in levels should give more accurate forecasts than the DVAR. This turns out to be the case only at short horizons. According to Christoffersen and Diebold (1998), the problem is that the ECM imposes both integration (unit roots) and cointegration, while the VAR in levels imposes neither. When the DGP contains unit roots (e.g. cointegrating vectors, differenced variables), and a VAR in levels is estimated, the estimation errors amplify over time. The VAR in levels is a poor forecaster because it fails to impose integration (unit roots). Again, the simulations with cointegrated DGPs support this theory.

If all variables are $I(1)$ and there are no groups of variables that cointegrate, estimation of a VAR in differences should give a better result and more accurate forecasts than any less-restricted model. This is the DGP for Lin and Tsay's (1996) model 5 (LT5) and again simulation evidence supports the theory. The interesting case is Lin and Tsay's model 3 which is stationary but contains two near-unit roots that a test would be unlikely to distinguish from unit roots. There is little to choose between estimation of a VAR in levels and estimation as an ECM, although Christoffersen and Diebold's arguments suggest that the ECM should be superior. Simulation studies where theory and simulation evidence conflict are the pioneering study by Engle and Yoo (1987) and the study with seasonal unit roots and cointegration (which lead to many potential models) by Lyhagen and Löf (2003).

We turn now to a more detailed comparison of the studies. Sometimes, a relatively minor difference in parameterization produces a different conclusion, as with the first two studies considered. Engle and Yoo (1987) used a two-variable VAR with one lag, no intercept and one cointegrating vector. Imposing the cointegration restriction instead of estimation in levels gives better forecasts, at long horizons, although not at shorter (up to six steps ahead). Clements and Hendry (1995) repeated the experiment, though with somewhat different parameter values that resulted in slower speed of adjustment (or quantitatively less 'error correcting'). Using the trace of the mean squared forecast error (TMSFE), the same measure as in Engle and Yoo (1987), they found that the Engle–Granger estimation of the VECM was more accurate than estimation of the VAR in levels for all horizons ($h = 1, 5, 10, 20$). The superiority of the VECM over estimation in levels is unchanged when using Clements and Hendry's preferred measure, the determinant of the second-moment matrix of stacked forecast errors, GFESM. But the choice of accuracy criterion can matter. With their preferred measure, estimation of variables in differences is worst at all horizons. These results are entirely consistent with theory. They are subject to the same caveats about robustness under model complexity, unknown structure and structural breaks as for Diebold and Kilian (2000).

Reinsel and Ahn (1992) used a larger VAR, with four variables and two lags and imposed two unit roots. They established critical values for a likelihood-ratio test for the number of unit roots (or equivalently the number of cointegrating vectors) and found that the test had good size properties. They estimated models with one to four unit roots and obtained TMSFE for one- to 25-step-ahead out-of-sample forecasts. Specifying fewer unit roots is harmless at short horizons ($h = 1, 2$) and damaging at long ones, while imposing more unit roots has exactly the opposite effect. In fact, imposing three unit roots gives the lowest TMSFE for $h > 12$. The TMSFE increases by a factor of 100 from one to 25 steps ahead.

Lin and Tsay (1996) used the same set-up as Reinsel and Ahn (1992) but considered more DGPs. These included: clearly stationary, with two near-unit roots, with two unit roots (and therefore two cointegrating vectors), and with four unit roots. With a slightly different parameterization, TMSFEs increased by a factor of 10 over the one- to 25-step-ahead horizons compared with a factor of 100 for Reinsel and Ahn's corresponding DGP. Qualitatively, results were the same. Out-of-sample forecast horizons ranged from one to 60 steps ahead. For each DGP, Lin and Tsay (1996) also estimated models with zero to four unit roots. For the clearly stationary models, imposing any unit roots worsens the forecast at any horizon, although with characteristic roots of 0.95, the damage is slight at short horizons and little affected by the number of unit roots imposed. When the largest characteristic roots are 0.99, imposing unit roots helps, with some benefit to imposing more rather than less. In the interesting middle case with two unit roots, most accurate forecasts result from imposing the correct number of unit roots. Specifying fewer unit roots is harmless at short horizons and damaging at long ones, while imposing more unit roots has exactly the opposite effect. When the DGP contains four unit roots, again, specifying fewer is harmless at short horizons and increasingly damaging at longer horizons. There is also progressively less accuracy as fewer unit roots are imposed. Overall, the recommendation seems to be: avoid imposing unit roots (i.e. estimate in levels) for horizons shorter than about six periods, otherwise, err on the side of extra unit roots. This finding is for a constant DGP, with large sample sizes (400) and so should not be affected by structural breaks or data-mining problems.

Similar results hold with seasonal data, where the possibilities for unit roots and cointegrating vectors are more complicated. Lyhagen and Löf (2003) examined seven different bivariate DGPs simulating quarterly data where one of the variables contained various combinations of unit roots at the zero, annual and biannual intervals. They concluded that in every case better out-of-sample forecasts resulted from cointegration models where both variables are transformed by annual differencing (imposing four unit roots) than by following the results of seasonal unit-root tests. A VAR in annual differences also forecasts more accurately than test-based cointegration models.

Results hold generally when the DGP is in logarithms and estimation is done (incorrectly) using variables in natural numbers and vice versa. But making the correct transformation (e.g. estimating with variables in logs when DGP is in logs) has a much bigger impact on forecast accuracy than imposing the correct number of unit roots. At least for the one-step ahead forecast they considered, Chao, Corradi and Swanson (2001) found that estimation in differences gives more accurate forecasts than estimation in levels (i.e. imposing too many unit roots) when the DGP is in logs and estimation is in natural numbers. To the extent that the DGP in logs mimics the non-linearities and breaks found in real-world data, this finding supports the idea that over-restricting the model gives it robustness.

There are some limitations to these Monte Carlo studies. The DGPs are typically very simple. Both DGP and estimating equation have a fixed-parameter structure (no time variation or structural breaks). And in many of the studies correct lag order is assumed, not tested. Also, unit-root and cointegration tests are not performed. The studies answer the question: 'Given a system with cointegrated variables, are more accurate forecasts achieved by imposing more, the correct number, or fewer than the correct number of restrictions?' Except for Diebold and Kilian (2000) they do not directly answer the question: Will unit root and cointegration tests reliably tell you the correct specification?

V. Empirical evidence

Despite the voluminous literature on cointegration and unit-root testing (growing steadily from 1983 and both peaking, at least temporarily, in 1999 in the Econlit database at 135 and 106 items respectively) and the accolade of the 2003 Nobel Prize, the empirical evidence of their utility is certainly not voluminous. We focus on forecast performance, where utility is easier to measure than, for example, in policy analysis. Table 2 includes only studies that compare the out-of-sample performance of models, published in the years 1984–2003. Comparisons are mostly one-step-ahead forecasts; few studies are available for longer horizons. They are further limited by the failure of most

TABLE 2

Number of series for which one strategy is better than another, by out-of-sample forecast accuracy of the resulting models, measured as RMSE, lead times not specified but mostly one step ahead

	Number of cointegrating vectors found or assumed														
	0			*1*			*2*			*3*			*Tota1 1–3*		
Strategies	*First best*	*Second best*	*Total*	*First best*	*Second best*	*Total*	*First best*	*Second best*	*Total*	*First best*	*Second best*	*Total*	*First best*	*Second best*	*Total*
Unrestricted vs. restricted lag order (1 vs. 1a)													7	27	35
Levels vs. differences (1 vs. 2)	0	8	8	6	6	12	0	1	1	2	5	7	8	12	20
VAR in levels vs. ECM from theory (1 vs. 3)				1	5	6							1	5	6
Unrestricted VAR in levels vs. pretest (1 vs. 4b)	2	1	5	18	7	26	1	6	7	3	8	11	22	21	44
Restricted VAR in levels vs. pretest (1a vs. 4b)				6	3	9							6	3	9
Total 1 vs. 3 and 4	2	1	5	25	15	41	1	6	7	3	8	11	29	29	59
VAR in differences vs. ECM from theory (2 vs. 3)				2	4	6	12	31	43				14	35	49
VAR in differences vs. pretest (2 vs. 4b)	8	2	10	9	19	29	0	1	1	6	1	7	15	21	37
Total 2 vs. 3 and 4	8	2	10	11	22	34	12	32	44	6	1	7	29	55	85

Note: Where the number in the ‘total’ column exceeds the sum of the preceding pair of columns, the difference is the number of series where the two strategies are equally accurate.

Simulation evidence excluded. The studies that comprise these results and their individual codings are listed in Appendix A available at http://www.umass.edu/resec/allen/obesapp.pdf.

studies to consider error measures beyond root mean squared error (RMSE), whereas comparative model performance is known to depend on the error measures used (e.g. Clements and Hendry, 1993). The table excludes comparisons with Bayesian VARs in levels or differences or Bayesian ECMs as falling outside the LSEM methodological framework. A distinction is made between unrestricted and restricted VARs since the difference does appear to matter. ‘Unrestricted’ means that all variables in all equations have the same lag length. The length is not necessarily chosen arbitrarily. In four of the seven studies (23 of 35 series) likelihood-ratio tests were used to reduce the lag lengths from the initial choice. Even so, as shown in the first line of Table 2, further simplification leads to better forecasts, supporting strategy 1a.

A model with variables entered in first differences (strategy 2) tends to give more accurate forecasts than the same model with the variables in levels (strategy 1), 20 series vs. eight, supporting strategy 2 (always difference). When there are no cointegrating vectors, estimation in differences is more efficient than estimation in levels or as an ECM. In the presence of structural breaks, estimation in differences is also likely to give more accurate forecasts when there are cointegrating vectors.

Estimating a VAR in levels (strategy 1) vs. performing a unit-root pretest and following its conclusion (usually to a cointegration test and an ECM) appears to have little impact; the strategies are essentially tied – a surprising result given the emphasis on simplification in the forecasting literature. Whilst this could in principle be due to sample-size effects (as the unconstrained-model estimates converge on the correctly specified ECM in large samples) such an explanation is unlikely for the sample sizes used in empirical work.

For most economic data, the commonest test result is to find one cointegrating vector. It is especially surprising that the derived ECM is less accurate than the less restrictive VAR in levels, a finding in conflict with both the principle of parsimony and results from Monte Carlo studies.

On the contrary, the bottom panel of Table 2 is much clearer in supporting the view that specifying a VAR in differences (strategy 2) is a bad approach compared with arriving at a (less restrictive) ECM either from theory or through testing. The VAR in differences comes out best when it should: when there are no cointegrating vectors, and rather surprisingly, in the less restrictive case when there are three.

We also attempted to understand behaviour at longer horizons by taking all the studies that reported out-of-sample forecasts for longer horizons and noting how estimation methods compared. The resulting table, Table 3, is directly comparable with Table 1. Empirical findings are, as expected, more diverse than simulation results, although some confirmations are found. Systems with no cointegrating vectors should be estimated as DVARs. The major exception of 43 series where DVAR is always worse and worsening

TABLE 3

Out-of-sample forecast accuracy results from empirical studies

	Forecast error increases with horizon		*Total*		*Series by number of cointegrating vectors*			
	Without limit	*To a ceiling*	*Series*	*Studies*	*0*	*1*	*2*	*3*
ECM vs. VAR is:								
Always better								
Improving	22 (3)	2 (1)	24	4	10	7	6	1
Worsening								
No pattern	3 (2)	2 (2)	5	4		3	1	1
Always worse								
Improving		1 (1)	1	1		1		
Worsening	3 (2)	9 (1)	12	3	5	1		6
No pattern		2 (2)	2	2		1		1
No pattern or varies								
Improving	4 (3)	1 (1)	5	4		2		3
Worsening	10 (3)	3 (2)	13	5	3	7	1	2
No pattern	3 (2)		3	2	1		3	
DVAR vs. VAR is:								
Always better								
Improving	23 (4)		23	4	11	5	5	2
Worsening	1 (1)		1	1				1
No pattern	4 (2)	1 (1)	5	3	3	1		1
Always worse								
Improving	1 (1)	3 (1)	4	2		4		
Worsening	2 (2)	4 (1)	6	3		1		3
No pattern	1 (1)		1	1				1
No pattern or varies								
Improving	3 (3)		3	3		1		2
Worsening		10 (4)	10	4	5	4		1
No pattern		1 (1)	1	1				1
DVAR vs. ECM is:								
Always better								
Improving	18 (7)	6 (3)	24	10	12	5		7
Worsening	1 (1)	1 (1)	2	2		2		
No pattern	1 (1)		1	1				3
Always worse								
Improving	2 (2)	1 (1)	3	3		2		1
Worsening	48 (3)	7 (2)	55	5	43	12		
No pattern	2 (2)	55 (1)	57	3				2*
No pattern or varies								
Improving	2 (2)	1 (1)	3	3		3		
Worsening	8 (4)	1 (1)	9	5	2	6		1
No pattern	15 (6)	6 (2)	21	8	1	14	5	1

Notes: VAR vector autoregression in levels, ECM unit-root restrictions imposed, DVAR unit-root restrictions imposed on all variables (first differenced).

Forecast accuracy usually measured as TMSFE, the trace of the mean squared forecast error.

'Always better' means that the forecast error from the first model is less than the forecast error from the second model for all horizons, opposite for 'always worse' and 'no pattern or varies' indicates either a single or multiple switches between better and worse. 'Improving' means that the forecast error from the first model consistently becomes smaller relative to the accuracy or the second model as the forecast horizon increases, 'worsening' means the opposite and 'no pattern' means no consistent direction of change.

*Number unknown for SS variables from one study.

Details are listed in Appendix A available at http://www.umass.edu/resec/allen/obesapp.pdf.

compared with ECM is from three studies, one of which reported an average of 38 variables. That study found that employment, hourly wages and hours worked per week were found to have zero cointegrating vectors in 38 of the 50 industries studied. Forecasts of each variable at each horizon were multiplied together (to give 'payroll') and the 38-industry average reported at each horizon for ECM and DVAR methods.

Results on series from systems where one cointegrating vector was detected, which represents the commonest situation, are particularly diffuse. When ECM and VAR models are compared, only seven of 22 series fall in the anticipated 'always better and improving' category, and in the DVAR and ECM comparison only 12 of 44 series fall in the anticipated 'always worse and worsening' category. Series from systems where three cointegrating vectors were detected are relatively unrestricted compared with the VAR in levels and so the VAR might be expected to perform strongly. The number of series falling in the anticipated 'always worse and worsening' is larger than the number falling in other cells except for the DVAR and ECM comparison.

Examination of the number of observations within sample, or the length of the most distant horizon, fails to show any pattern that explains the many discrepancies. Different series from the same system (in the same study) show up in a variety of categories. A substantial proportion of the comparisons show up in the 'no pattern or varies' category. Given that the empirical results are so scattered, it is hard to make a strong case for unit root and cointegration testing compared with either assuming cointegration on theoretical grounds or simply estimating a DVAR.

VI. Conclusions

Each of the strategies identified has advantages and disadvantages, either theoretical or empirical. The differences arise from the benefits of imposing restrictions when they hold true out of sample compared to the costs if they fail.

Compared with other specifications, strategy 1, an 'unrestricted VAR in levels' avoids throwing away information (Sims, 1980). Even if the true model is a VAR in differences, hypothesis tests based on a VAR in levels will have the same asymptotic distribution as if the correct model had been used. However, it may be overparameterized and give correspondingly bad forecasts.

But the initial unrestricted model, like all the alternative approaches to model specification, is no more immune from failing mis-specification tests (wrong choice of variables, poor autoregressive approximation to the true DGP, etc.; Harvey, 1997). It also responds slowly to structural breaks. Comparing the unrestricted VAR with its restricted cousin, the simpler model, following similar conclusions from univariate comparisons (Fildes and Ord, 2002), proves the more accurate as Table 2 shows.

The ‘restricted VAR in levels’ specification, strategy 1a, also may ignore data-congruent restrictions. These derive from long-run equilibrium relationships (cointegration) that would lead to alternative, even simpler, model structures through ‘post-testing’ strategy 1b. We found no studies that compared strategies 1 and 1b directly, only comparisons of strategies 1 and ‘pretesting’, 4b. But the empirical evidence of Table 2 offers little support for the view that this leads to a significant improvement in forecasting. However, the ‘post-testing’ strategy rarely leads to a VAR in differences, so this remains the preferred strategy as ECMs prove considerably better than VARs automatically specified in differences (strategy 2). Strategy 2 also suffers from the problem that if the variables are already stationary, differencing induces a moving average term into the equation (though how important such a mis-specification is when estimating the model is an empirical question). Although Monte Carlo studies have shown that an advantage of specifying a VAR in differences is that it is robust if there are structural breaks, we found no empirical evidence to support the simulation comparisons.

Strategy 3 relies on other means of model specification (such as cointegration relations imposed from theory). A general model that is theoretically consistent allows specification testing for nested special cases (e.g. a time-varying parameter model that allows for fixed parameters as special cases). To support his argument for this strategy (Harvey, 1997, p. 197), he focuses on the near impossibility of identifying the appropriate order of integration. Empirical evidence, summarized in Allen and Fildes (2001), suggests that time-varying parameter or state–space models developed with this strategy forecast better than models developed by other methods.

The final arm of Figure 1 relies on pre-tests. The use of unit root pre-tests to ensure a model specified in stationary variables has the single advantage of achieving a constrained model. But the tests have low power and might lead to erroneous conclusions about the existence of unit roots and cointegrating vectors, resulting in a misspecified over-constrained model. Nor are the constraints necessarily appropriate (consider an ECM model with one cointegrating vector).

Assembling the evidence presented so far suggests that the strategies of never testing for unit roots and cointegration (strategies 1–3) are inferior to testing (strategy 4), even though the tests have admittedly low power. It does not much matter whether unit-root and cointegration testing is conducted on the variables before a model is specified or as part of the reduction of an acceptable general model. What does matter is that the unit-root pretest not be used to establish a set of $I(0)$ variables (by differencing where necessary) and these variables be entered into a VAR (strategy 4a). This is a form of restricted VAR that should have desirable properties, as all variables are in stationary form, but fails to make use of the valuable information contained in a cointegrating relationship.

Ideally, tests will give information on the correct number and form of restrictions, so that unit-root and cointegration restrictions imposed on the final model are those that best describe the DGP. According to simulation evidence, imposing one more restriction than the correct number is harmless. This finding is likely to hold with real data where structural breaks are commonplace and imposing additional unit root restrictions will therefore make the final model more robust to breaks.

To conclude, the primary aim of this paper has been to establish some clear principles of model specification for the purposes of forecasting within the LSE methodological framework. Empirical evidence we have collected shows that test-based strategies to reduce the general model in levels to a constrained model with either ECM or differenced variables are beneficial and lead to improved forecasting accuracy. The results are generally in accord with the predictions derived from cointegration theory. Within the LSE methodology we can therefore claim to have established a principle that forecasters should build into their modelling: start with a general model and use unit-root and cointegration tests (strategies 1b and 4b). Other studies have established the value of reducing the length of lag on all variables to a well specified reduced model and because of the great reduction in number of parameters to estimate, such reduction should occur first.

The principle just established is unambiguous but conflicts with much of the advice given in the textbooks. Nor does it appear to be uniformly true. Further investigation should therefore lead towards a clearer specification of the conditions under which it holds. The empirical evidence is not overwhelming and, in some circumstances, is in conflict with the theoretical predictions, e.g. where an ECM is outperformed by a model specified in levels, even though a cointegrating vector has been detected. Of course the number of studies involved is small and the results stochastic. Structural breaks we know to be a factor that leads to the better performance of differenced model specifications and this may explain the observed results. The principle is much in need of refinement; e.g. Phillips (2003) essentially states a principle when he says that prediction from a model with misspecified trend may not be that serious provided we make an effort to keep the model on track by using intercept adjustments. As the evidence slowly accumulates on these strategies, with more careful testing of structural breaks as an automatic aspect of specification testing, both in and out of sample (and a more detailed consideration of the resulting forecast error statistics), we can expect this anomaly to be clarified.

A subsidiary aspect of the paper has been to examine the role of algorithmic vs. tacit knowledge when specifying forecasting models. Kennedy (2002) cites the complaints of several well-known econometricians: relevant here is 'we teach what we know, not the applied econometrics needed for analysis of messy data' and 'writing down rules to guide data analysis (and when to ignore them) is hard, because so much of data analysis is subjective, subtle and a tacit skill'. This unsatisfactory state of affairs is changing, if slowly. Computer algorithms, or expert systems, such as GETS and ModelBuilder require explicit rules. Extending the computer code and measuring the effect on outcomes shows the impact of well-defined rules. Perhaps a parallel situation is in the estimation of ARIMA models. As proposed by Box and Jenkins, considerable experience, judgment and time were called for to identify, estimate and evaluate such models. Today, computer algorithms routinely produce models as good as or better than the experts, in a fraction of the time. Econometric analysis is considerably more involved, but the same progression should be possible.

The analysis of a wide range of empirical evidence carefully coded has shown its worth, when interpreted through econometric theory. Potential anomalies have been identified suggesting areas of future research. Such an approach has the potential for driving downward Pagan's 'uncomfortably high art-to-science ratio'.

References

Allen, P. G. and Fildes, R. (2001). 'Econometric forecasting', Chapter 11, in Armstrong J. S. (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer Academic Press, Norwell, MA, pp. 303–362.

Chao, J. C., Corradi, V. and Swanson, N. R. (2001). 'Data transformation and forecasting in models with unit roots and cointegration', *Annals of Economics and Finance*, Vol. 2, pp. 59–76.

Christoffersen, P. F. and Diebold, F. X. (1998). 'Cointegration and long-horizon forecasting', *Journal of Business and Economic Statistics*, Vol. 16, pp. 450–458.

Clements, M. P. and Hendry, D. F. (1993). 'On the limitations of comparing mean square forecast errors, with discussion', *Journal of Forecasting*, Vol. 12, pp. 617–637.

Clements, M. P. and Hendry, D. F. (1995). 'Forecasting in cointegrated systems', *Journal of Applied Econometrics*, Vol. 10, pp. 127–146.

Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*, Cambridge University Press, Cambridge.

Darnell, A. C. and Evans, J. L. (1990). *The Limits of Econometrics*, Elgar, Aldershot.

Dharmapala, D. and McAleer, M. (1996). 'Econometric methodology and the philosophy of science', *Journal of Statistical Planning and Inference*, Vol. 49, pp. 9–37.

Dickey, D. A. and Fuller, W. A. (1979). 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association*, Vol. 74, pp. 427–431.

Diebold, F. X. (1998). *Elements of Forecasting*, South-Western College Publishing, Cincinnati, OH.

Diebold, F. X. and Kilian, L. (2000). 'Unit root tests are useful for selecting forecasting models', *Journal of Business and Economic Statistics*, Vol. 18, pp. 265–273.

Elder, J. and Kennedy, P. E. (2001). 'Testing for unit roots: what should students be taught?', *Journal of Economic Education*, Vol. 32, pp. 137–146.

Engle, R. F. and Yoo, B. S. (1987). 'Forecasting and testing in co-integrated systems', *Journal of Econometrics*, Vol. 35, pp. 143–159.

Fildes, R. and Ord, J. K. (2002). 'Forecasting competitions – their role in improving forecasting practice and research', Chapter 15, in Clements M. and Hendry D. (eds), *A Companion to Economic Forecasting*, Blackwell Publishing, Oxford, pp. 322–353.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A. (1997). 'Trends, cycles and autoregressions', *Economic Journal*, Vol. 107, pp. 192–201.

Hendry, D. F. (1995). *Dynamic Econometrics*, Oxford University Press, Oxford.

Hendry, D. F. (1997). 'The econometrics of macroeconomic forecasting', *Economic Journal*, Vol. 107, pp. 1330–1357.

Hendry, D. F. (1999). 'An econometric analysis of US food expenditure', Chapter 17, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 341–361.

Hendry, D. F. (2002). 'Applied econometrics without sinning', *Journal of Economic Surveys*, Vol. 16, pp. 591–604.

Hendry, D. F. and Krolzig, H.-M. (2003a). 'New developments in automatic general-to-specific modeling', Chapter 16, in Stigum B. P. (ed.), *Econometrics and the Philosophy of Economics: Theory-Data Confrontations in Economics*, Princeton University Press, Princeton, NJ, pp. 379–419.

Hendry, D. F. and Krolzig, H.-M. (2003b). *The Properties of Automatic Gets Modeling*, Nuffield College Economics Working Papers 2003-W14. http://www.nuff.ox.ac.uk/economics/papers/2003/W14/dfhhmk03a.pdf.

Holden, K. (1995). 'Vector autoregression modeling and forecasting', *Journal of Forecasting*, Vol. 14, pp. 159–166.

Hoover, K. D. and Perez, S. J. (1999). 'Data mining reconsidered: encompassing and the general-to-specific approach to specification searches', *Econometrics Journal*, Vol. 2, pp. 167–191.

Kennedy, P. E. (2002). 'Sinning in the basement: what are the rules? The ten commandments of applied econometrics', *Journal of Economic Surveys*, Vol. 16, pp. 569–589.

Keuzenkamp, H. A. and McAleer, M. (1995). 'Simplicity, scientific inference and econometric modelling', *Economic Journal*, Vol. 105, pp. 1–21.

Kurcewicz, M. (2002). 'ModelBuilder – an automated general-to-specific modeling tool', in Haerdle W. and Roenz B. (eds), *Compstat 2002 – Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 467–472. Also available at http://www.quantlet.de/scripts/compstat2002_wh/paper/full/M_07_kurcewicz.pdf.

Leamer, E. E. (1999). 'Revisiting Tobin's 1950 study of food expenditure', Chapter 5, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 123–152.

Lin, J.-L. and Tsay, R. S. (1996). 'Co-integration constraint and forecasting: an empirical examination', *Journal of Applied Econometrics*, Vol. 11, pp. 519–538.

Lyhagen, J. and Löf, M. (2003). 'On seasonal error correction when the processes include different numbers of unit roots', *Journal of Forecasting*, Vol. 22, pp. 377–389.

Maddala, G. S. and Kim, I.-M. (1998). *Unit roots, Cointegration, and Structural Change*, Cambridge University Press, Cambridge.

Magnus, J. R. and Morgan, M. S. (1999a). *Methodology and Tacit Knowledge*, Wiley, Chichester.

Magnus, J. R. and Morgan, M. S. (1999b). 'Lessons from the tacit knowledge experiment', Chapter 20, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 375–381.

Magnus, J. R. and Morgan, M. S. (1999c). 'Lessons from the field trial experiment', Chapter 14, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 301–307.

Owen, P. D. (2003). 'General-to-specific modelling using PcGets', *Journal of Economic Surveys*, Vol. 17, pp. 609–628.

Pagan, A. R. (1987). 'Three econometric methodologies: a critical appraisal', *Journal of Economic Surveys*, Vol. 1, pp. 3–24.

Pagan, A. R. (1999). 'The Tilburg experiments: impressions of a drop-out', Chapter 19, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 369–374.

Phillips, P. C. B. (2003). 'Laws and limits of econometrics', *Economic Journal*, Vol. 113, pp. C26–C52.

Reinsel, G. C. and Ahn, S. K. (1992). 'Vector autoregressive models with unit roots and reduced rank structure: estimation, likelihood ratio test and forecasting', *Journal of Time Series Analysis*, Vol. 13, pp. 353–375.

Siegert, W. K. (1999). 'An application of three econometric methodologies to the estimation of income elasticity of food demand', Chapter 16, in Magnus J. R. and Morgan M. S. (eds), *Methodology and Tacit Knowledge*, Wiley, Chichester, pp. 315–340.

Sims, C. A. (1980). 'Macroeconomics and reality', *Econometrica*, Vol. 48, pp. 1–48.

Stock, J. H. and Watson, M. K. (2003). *Introduction to Econometrics*, Addison-Wesley, Boston, MA.