doi: 10.1111/j.1467-6419.2007.00518.x

DIRECT MULTI-STEP ESTIMATION AND FORECASTING

Guillaume Chevillon

ESSEC, Paris, and University of Oxford

Abstract. This paper surveys the literature on multi-step forecasting when the model or the estimation method focuses *directly* on the link between the forecast origin and the horizon of interest. Among diverse contributions, we show how the current consensual concepts have emerged. We present an exhaustive overview of the existing results, including a conclusive review of the circumstances favourable to *direct* multi-step forecasting, namely different forms of non-stationarity and appropriate model design. We also provide a unifying framework which allows us to analyse the sources of forecast errors and hence of accuracy improvements from *direct* over *iterated* multi-step forecasting.

Keywords. Multi-step forecasting; Adaptive estimation; Varying horizon; Structural breaks; Non-stationarity

1. Introduction

Economic forecasting is a task distinct from that of modelling because it has been shown (see *inter alia* Clements and Hendry, 1999; Allen and Fildes, 2001; Fildes and Stekler, 2002) that causal models do not necessarily forecast better that non-causal alternatives. Rather, causal models often suffer forecast failure and many adjustment techniques have been developed such as intercept corrections (see for example Clements and Hendry, 1998a). Other routes include the use of non-congruent models or 'naive' formulations – such as constant growth or random walk hypotheses – which often enhance accuracy owing to their robustness to instability (structural breaks, regime change, economic policy shifts, technological discoveries...) which generate misspecification in the economic models.

When a modeller wishes to produce forecasts at several horizons, an intuitively appealing idea, 'direct multi-step estimation' (DMS), consists in matching model design with the criterion used for its evaluation. Hence, DMS directly minimizes the desired multi-step function of the in-sample errors and offers a potential way to avoid some of the aforementioned difficulties. By contrast, the standard procedure uses one-step estimation – via minimizing the squares of the in-sample one-step-ahead residuals – from which multi-step forecasts are obtained by 'iterated multi-step' (denoted here by IMS). One intuition behind DMS is that a model which is misspecified for the data-generating process (DGP) need not be a satisfactory forecasting device. However, misspecification is insufficient: predictors like constant growth are misspecified but robust. Here, the desired robustness is to misspecification of the model dynamics or *vis-à-vis* unnoticed parameter change. Among model misspecifications which might sustain DMS, unnoticed unit roots stand out; neglected serial correlation of the disturbances also provide a rationale at short horizons. In stationary processes, DMS could enhance forecast accuracy, but gains fade rapidly as the horizon increases.

The idea of multi-step estimation has a long history and its developments have followed many paths. Two main DMS approaches have been studied. First, for the parametric technique, the same model parameters are estimated via minimizing distinct horizon-dependent criteria; the techniques used in this case are most often non-linear, and the model may or may not be assumed misspecified. By contrast, non-parametric DMS focuses on the parameters of a different – misspecified beyond the first step – model at each horizon.

The purpose of this paper is to review the main contributions to the multi-step forecasting literature and to show how they arose in order to provide a unifying treatment of the many distinct existing results. We first need to explain what we mean by *direct* multi-step forecasting but its definition has emerged only progressively and we think it preferable not to define it at this stage but only after reviewing the various contributions: this concept has historically served as an umbrella for different approaches and only one has proved useful in forecasting. Let us clarify for now that the traditional estimation method consists of estimating, for a vector variable $\mathbf{x}_t$, the equation relating it to its past and, potentially, to additional variables. If we denote by $\mathcal{F}_t$ the sigma field representing the information available at time t, the traditional method seeks to model and estimate how $\mathbf{x}_t$ is generated given $\mathcal{F}_{t-1}$, or $\mathbf{x}_t|\mathcal{F}_{t-1}$, so as to produce an equation such that

$$\mathbf{x}_t = \widehat{\mathbf{f}}(\mathcal{F}_{t-1}) \quad \text{for } t \leq T \tag{1}$$

From a date T, when $\mathcal{F}_T$ is available, it is therefore possible, using the estimated (1), to generate a forecast for $T + 1$, namely

$$\widehat{\mathbf{x}}_{T+1|T} = \widehat{\mathbf{f}}(\mathcal{F}_T)$$

assuming that the intertemporal link between $\mathbf{x}_t$ and $\mathcal{F}_{t-1}$ will remain valid in the future. When $\mathcal{F}_T$ is generated only by $\{\mathbf{x}_t\}_{t\leq T}$, the same assumption about $\mathbf{x}_{T+h}$ and $\mathcal{F}_{T+h-1}$, for $h > 1$, leads to replacing $\mathcal{F}_{T+1}$ by $\widehat{\mathcal{F}}_{T+1}$ which we regard as pseudo information relating to $\{\ldots, \mathbf{x}_T, \widehat{\mathbf{x}}_{T+1|T}\}$ where $\widehat{\mathbf{x}}_{T+1|T}$ is assumed to be authentic information (in reality $\widehat{\mathcal{F}}_{T+1} = \mathcal{F}_T$), so that we produce

$$\widehat{\mathbf{x}}_{T+2|T} = \widehat{\mathbf{f}}(\widehat{\mathcal{F}}_{T+1})$$

and so on, for higher forecast horizons. We define the resulting forecasts as iterated multi-step or IMS.

By contrast, an alternative method consists of directly estimating the relationship of interest at the hth horizon, namely $\mathbf{x}_t|\mathcal{F}_{t-h}$, so that a DMS forecast is generated by

$$\widetilde{\mathbf{x}}_{T+h|T} = \widetilde{\mathbf{k}}_h(\mathcal{F}_T)$$

Care must be paid to the terms used: the one-step (1S) parameter estimates (which coincide for both IMS and DMS at $h = 1$) are those obtained for $\widehat{\mathbf{f}}(\cdot)$; they imply some IMS counterparts by combination and powering up. By contrast, the DMS parameters are *directly* estimated. Thus, the main distinction between the two methods is that IMS forecasting necessitates only one estimation procedure but the estimates are modified for each horizon, whereas DMS needs re-estimation for each h, but then such estimates are directly usable. We note, and will see below, that some authors define DMS as an estimation of the one-step parameters using a non-linear criterion based on $\mathbf{x}_t|\mathcal{F}_{t-h}$; this, seemingly, uncouples estimation and forecasting and does not correspond to our choice of definition, although both are related, which will lead us to consider this case too. We provide below the eight key steps of the progressive research which explain what the state of knowledge now is; each literature strain provides the opportunity for a discussion.

We organize our analysis as follows. Section 2 explains the first instances when it was suggested to resort to some dynamic, rather than one-step, estimation. The following section makes explicit the results regarding the inefficiency from using a multi-step procedure to estimate the one-step-ahead parameters of a well-specified model (which we call parametric DMS) and we show the need for model misspecification in Section 4. The general theoretical results regarding multi-step forecasting are presented in Section 5. We turn to non-parametric estimation and the design of forecasting models in Section 6; robustness towards breaks is analysed in Section 7 and Section 8 presents the features that favour the use of DMS. Section 9 concludes the review of literature by presenting recent work on testing for improved multi-step forecast accuracy using direct methods. After reviewing all the progress made in the literature and the many aspects covered, we finally present our analysis of the general framework appropriate for the analysis of direct multi-step estimation and forecasting in Section 10 and show that it explains how to interpret the results found in the existing literature. A conclusion provides paths for future research.

2. Early Suggestions: Estimate the Dynamic 'Solution Path'

The first instance when some dynamic estimation was suggested is found in Cox (1961) who compares the mean square forecast errors (MSFEs) from an exponentially weighted moving average (EWMA) model and an autoregressive (AR(1)) model with an intercept when the true DGP is either AR or autoregressive moving average (ARMA) with an intercept. Cox shows that, if the mean of the process to be forecast is allowed to shift, the parameters of the prediction model should depend on the forecasting horizon so that robustness can be achieved. He suggests combining the EWMA and the AR forecasting techniques with weights which vary with the horizon.

At the turn of the 1970s, several authors started focusing on the difficulties in estimating dynamic models. Their concern was that of predetermined variables and their interest lay in the design of estimation techniques which take full advantage of the dynamic structure of the series.

Klein (1971) suggested a multi-step estimator which minimizes the 'solution path' as mentioned in Haavelmo (1940). His idea was that in general if the DGP follows an AR(1) (which can readily be extended to include more lags or exogenous variables),

$$y_t = \alpha y_{t-1} + \epsilon_t \quad \text{for } t = 1, \ldots, T, \text{ and } |\alpha| < 1$$

and one wishes to obtain forecasts of $y_{T+h} = \alpha^h y_T + \sum_{i=0}^{h-1} \alpha^i \epsilon_{T+h-i}$, for $h = 1, \ldots, H$, then least squares estimation of the model leads to minimizing the criterion function

$$\sum_{h=1}^{H} \sum_{t=1}^{T-h} \left(\sum_{i=0}^{h-1} \alpha^i \epsilon_{t+h-i} \right)^2 = \sum_{h=1}^{H} \sum_{t=1}^{T-h} \left(y_{t+h} - \alpha^h y_t \right)^2$$

with respect to the coefficient α. In a simulation experiment, Klein let several parameters vary and his findings were that (1) multi-step methods seem to perform better in smaller samples (here 50 versus 400), (2) adding a trendless exogenous variable seems to help DMS, but a trending variable does not, and (3) the initial observation does not affect the previous results. In applying this dynamic estimation method to the Wharton model, he found that he could reduce the mean average prediction error from 6.29% to 5.33% in two-step-ahead out-of-sample forecasting, when comparing it with an instrumental variables estimation with principal components.

Hartley (1972) studied the properties of the dynamic least squares estimator (DLS for him) suggested by Klein (1971) in the univariate AR(1) case. He showed that the new estimator was more robust to residual autocorrelation than ordinary least squares (OLS).

Assuming that the process can be written, for $t = 1, \ldots, T$, as

$$y_t = \alpha y_{t-1} + \epsilon_t \tag{2}$$

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \tag{3}$$

where y_0 is fixed, $\epsilon_0 = 0$, $\{u_t\}$ is an independently and identically distributed (i.i.d.) process whose elements have zero mean, variance σ^2 and finite third and fourth moments, $|\alpha| < 1$ and $|\rho| < 1$, Hartley showed that if the dynamic solution path

$$y_t = \alpha^t y_0 + \sum_{i=1}^{t} \alpha^{t-i} \epsilon_i$$

is estimated by generalized least squares (GLS), then it is the same as OLS when $\rho = 0$. Denoting the OLS and DLS estimators of α by, respectively, $\widehat{\alpha}$ and $\widetilde{\alpha}$, then

$$\widehat{\alpha} - \alpha \underset{T\to\infty}{\to} \frac{\rho(1-\alpha^2)}{1+\alpha\rho}$$

but $\widetilde{\alpha}$ does not converge unless it is assumed that $y_0 = \mathrm{O}_p(T^k)$, for $k > 0$, under which circumstance, if $\rho = 0$,

$$\widetilde{\alpha} - \alpha \underset{T\to\infty}{\to} \mathsf{N}\left[0, (1-\alpha^2)\frac{1+3\alpha^2+\alpha^4}{(1+\alpha^2)^2}\frac{\sigma^2}{y_0^2}\right]$$

so that the asymptotic variance of the DLS estimator is of order $1/T^{2k}$. The author showed that when $\rho \neq 0$ and $y_0 = \mathrm{O}_p(T^k)$, there exists a function $f(\cdot, \cdot)$ such that

$$\lim_{T\to\infty} \mathrm{Var}\,(\widetilde{\alpha} - \alpha) = f(\alpha, \rho)\,\frac{\sigma^2}{y_0^2}$$

Thus the DLS estimator is asymptotically unbiased, whereas OLS is not. Hartley noted that the requirement about the initial observation was satisfied even with very low k. Yet the variance could not be made arbitrarily small since σ should then increase with y_0. He also noticed that if the errors follow an MA(1) rather than an AR(1), then the DLS estimator is the maximum likelihood estimator.

Johnston *et al.* (1974) noticed that dynamic models which incorporate lagged values of the endogenous variable may lead to a contradiction between the assumptions made for estimation and for forecasting. Indeed, it has been common practice since the work by Mann and Wald (1943) to consider that the lags of the stationary endogenous variable can be asymptotically treated as ‘exogenous’, or predetermined. However, when formulating a forecast at several periods in the future, the intermediate lags – between the forecast origin and the period of the forecast – can no longer be seen to be predetermined and this aspect ought to be taken into consideration. They built their work on the previous results by Haavelmo (1944) who showed that for the case of a stationary AR(1) process with no drift

$$y_t = \alpha y_{t-1} + e_t \tag{4}$$

the optimal – in the sense of minimizing a quadratic loss function in e_{T+1} and e_{T+2} – prediction formulae for $T + 1$ and $T + 2$ from an end-of-sample forecast origin y_T are given by

$$y_{T+1} = \alpha y_T$$
$$y_{T+2} = \alpha^2 y_T$$

The significant aspect is that the maximum likelihood estimate of α is that of α^2 too.

Johnston *et al.* (1974) compared several forecasting methods for the AR(1) with different parameter values and applied their techniques to the Wharton model. Their idea was to compute estimators and resulting forecasts which incorporate the dynamic structure of the DGP. The hypothesis is that ‘systems using up to pth order generated lag values as instruments or regressors will perform best in p period point prediction’. In general, the DLS estimator for a model such as

$$\mathbf{A}(L)\,\mathbf{y}_t = \epsilon_t$$

where $\mathbf{A}(L)$ is a matrix polynomial and L the lag operator, is that which minimizes the criterion

$$\operatorname{tr} \sum_{t=1}^{T} \left(\mathbf{A}(L)^{-1} \epsilon_t\right) \left(\mathbf{A}(L)^{-1} \epsilon_t\right)'$$

In the univariate AR(1) case from (4), this is

$$\widetilde{\alpha} = \underset{\alpha}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - y_0 \alpha^t\right)^2 \tag{5}$$

The procedure used by the authors for actual minimization is a grid search. The results of Monte Carlo simulations with fixed or stochastic initial values and various stationary values of α show that the variance of the DLS estimator is higher than that of OLS when the initial value is stochastic, but lower for a fixed initial value. In terms of MSFE, their results are that for small samples (either 20 or 50 observations and forecast horizons, respectively, of 5 or 10 periods) DLS outperforms OLS when the initial value is fixed, but when the latter is stochastic, the forecast loss is lower for DLS only for very small samples.

The authors then used two-stage least squares estimators for the Wharton model. They matched the values of the lag used for the endogenous variable as an instrument and the horizon at which it is desired to forecast. Unfortunately, their results for out-of-sample prediction are somewhat inconclusive. Some gains are obtained at short horizons and seem to improve with the lead in the forecast for personal income and total consumption but not for the GNP deflator, investment in non-farm inventories and unemployment rate. The GNP deflator is the only variable for which the within-sample residual standard error and the post-sample root-MSFE are of the same magnitude. The latter is much larger than the former as regards the other variables.

2.1 *Discussion*

The concept of solution path is the first attempt to provide an estimation method which embodies the dynamics of the process. Unfortunately, its dependence on the initial observation makes it impractical since it can be strongly contaminated by measurement errors and it asymptotically relies on the artificial assumption that the initial observation increases with the sample size. Thus, this methodology is of little use for both stationary and integrated processes. Yet it paves the way for multi-step estimation where it is not the same initial observation which is used, but the same lag; thus leading, instead of (5), to

$$\widetilde{\alpha}_h = \underset{\alpha_h}{\operatorname{argmin}} \sum_{t=h}^{T} \left(y_t - y_{t-h} \alpha^h\right)^2$$

And, now, there is no longer any need for an exploding initial value. This augurs all the better for the use of DMS since the first simulation results by Johnston *et al.* seemed to confirm that such methods fare well in small samples, where the initial value need not be of magnitude different from the rest of the observations.

3. Inefficiency of DMS Estimation in a Well-specified Model

The first authors who analysed multi-step estimation techniques compared their asymptotic properties to those of other well-established methods when the model is well-specified for the stationary DGP.

Johnston (1974) analysed the forecasting properties of the multi-step estimator (dynamic estimator for him) suggested by Johnston *et al.* (1974) and compared them to those of a one-step-ahead estimator. His framework was that of a well-specified dynamic vector model with exogenous variables and mean zero errors

$$\mathbf{y}_t = \mathbf{z}_t \boldsymbol{\theta} + \epsilon_t \tag{6}$$

where $\mathbf{z}_t = (\underset{\frown}{\mathbf{y}}_t, \mathbf{y}_{t-1}, \mathbf{x}_t)$ and $\boldsymbol{\theta}' = (\mathbf{A}_0', \mathbf{A}_1', \mathbf{B}')$, with zero diagonal entries for $\mathbf{A}_0$. For an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, $\widehat{\mathbf{y}}_{T+h,h}$ is the forecast of $\mathbf{y}_{T+h}$ conditional on $\{\mathbf{y}_t\}_{t\leq T}$, $\{\mathbf{x}_t\}_{t\leq T+h}$ and $\widehat{\boldsymbol{\theta}}$, as obtained by some least squares technique. The user's loss function is assumed to be quadratic and given by

$$L(T, \underline{h}, \overline{h}) = \sum_{h=\underline{h}}^{\overline{h}} w_h (\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h,h}) \mathbf{Q} (\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h,h})' \tag{7}$$

where $\mathbf{Q}$ is a positive definite matrix weighting the importance of forecast errors across equations and the set of weights, $\{w_h\}$, expresses the relative importance of the forecasting horizons ($w_h \geq 0, \forall h$). The direct multi-step estimator is then defined as

$$\widehat{\boldsymbol{\theta}}_{\text{DMS}} = \underset{\widehat{\theta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T-\overline{h}} L(t, \underline{h}, \overline{h}) \right\} \tag{8}$$

It corresponds to the DLS whose distribution was derived by Hartley in the univariate case. And when the model is well specified – i.e. $\rho = 0$ in (3) – this estimator is asymptotically less efficient than the one-step OLS, thus being consistent with the claim in Haavelmo (1944) that, when the error loss function is quadratic, the rankings of the estimators in prediction and estimation efficiency match. Johnston's PhD thesis showed that, asymptotically, the 'optimal' estimator is invariant to the choice of – quadratic – prediction error loss function. Johnston maintained that, in practice, multi-step estimation can be justified if it is more efficient than an alternative computationally equivalent estimator. Yet, as the paper proves, the asymptotically most efficient – in terms of minimum variance – estimator from (8) is given by $\overline{h} = \underline{h} = 1$ (Johnston only considered the case where $\overline{h} = \underline{h}$, $\mathbf{Q} = \mathbf{I}$, $w_h = 1\ \forall h$, and where $\widehat{\boldsymbol{\theta}}_{\text{DMS}}$ is obtained by iterated minimization of (8), given an initial estimate $\widehat{\boldsymbol{\theta}}_0$ which provides $\widehat{\mathbf{z}}_t(\widehat{\boldsymbol{\theta}}_0)$, until convergence). The main result is that

$$\Sigma_{\overline{h}} - \Sigma_{\overline{h-1}} \geq 0 \tag{9}$$

where $\Sigma_{\overline{h}}$ is the asymptotic variance of the multi-step estimator (scaled by $\sqrt{T}$) for $\overline{h} = \underline{h}$. Thus the one-step OLS estimator has *minimum asymptotic variance* and is hence efficient. This author mentioned some unpublished simulation results which confirm this finding even in small samples. He noted, however, that small sample biases should be taken into account since they make the estimator variance and MSFE differ.

Kabaila (1981) was interested in comparing the asymptotic efficiency of the IMS and DMS estimators in general non-linear processes. His assumptions were that the process under consideration $\{y_t\}$ is strictly stationary and generated by

$$y_t = f(y_{t-1}, y_{t-2}, \ldots; \theta_0) + \epsilon_t$$

where the process $\{\epsilon_t\}$ is i.i.d. and its elements have zero expectation and variance σ^2, $\theta_0 \in \Theta \subset R^p$, and y_t is a measurable function of $\{\epsilon_t\}$. The dots in $f(\cdot)$ indicate that the initial values can be of any form and number. Obviously, $f(y_{t-1}, \ldots; \theta_0) = E[y_t | y_{t-1}, \ldots; \theta_0]$. Define, similarly,

$$g_k(y_{t-k}, y_{t-k-1}, \ldots; \theta_0) = E[y_t | y_{t-k}, \ldots; \theta_0] \quad \text{for } k > 1$$

The function $h_{k,t}(\cdot)$ is defined as that obtained by backward substitution of $f(\cdot)$ and ϵ_{t-j} for $j = 0, \ldots, k-1$, such that

$$g_k(y_{t-k}, y_{t-k-1}, \ldots; \theta_0) = E[h_{k,t}(\theta_0) | y_{t-k}, \ldots]$$

Kabaila (1981) made an additional assumption, namely that for all $\theta \in \Theta$

$$h_{k,t}(\theta) = \sum_{i=1}^{k-1} \epsilon_{t-i} U_{t,i}(\theta) + V_{t-k}(\theta)$$

where $U_{t,i}(\theta)$ $(i > k-1)$ is a function of θ, the ϵ_j and y_m for $j \leq t-i-1$ and $m \leq t-k$; $U_{t,k-1}(\theta)$ and $V_{t-k}(\theta)$ are functions of $y_{t-k}, y_{t-k-1}, \ldots$.

Let $\widehat{\theta}_T$ and $\widehat{\theta}_{k,T}$ denote minimizers – with respect to θ – of some approximations to the in-sample (for a sample of size T) sum of the squared residuals, respectively $y_t - f(y_{t-1}, y_{t-2} \ldots; \theta_0)$ and $y_t - g_k(y_{t-k}, y_{t-k-1}, \ldots; \theta_0)$. By approximation, it is meant that the initial values $y_{-1}, \ldots$ may not be known and this is reflected in the objective function.

Provided that the asymptotic variances of the estimators (which exist) are non-singular, Kabaila proved that the 1S estimator is efficient, as an estimator of θ.

3.1 *Discussion*

These authors were interested in comparing some parameter estimators which account for some of the dynamics of the process. This is one of the two strains of multi-step estimation and, unfortunately, brings no benefits. Here the *same* parameter is estimated by either one-step or h-step methods. It is simply the objective functions that differ, in so far as the h-step criterion is a non-linear composition of the one step. Indeed, in both cases, the h-step fitted values – $\widehat{\mathbf{y}}_{T+h,h}$ or $h_{k,t}(\theta)$ – are computed using the same model as that for 1S. Under these assumptions, the authors showed that one-step estimation is asymptotically more efficient than this type of DMS. The two contributions are thus essential, since they show that for DMS to provide any gains, one of the four following assumptions must be made: (1) the model is misspecified, (2) different models are used for 1S and DMS, (3) it is the implied (powered-up) multi-step parameters which are of interest, not the one-step estimated by a multi-step criterion, or (4) the gains are to be found in small samples. These assumptions are studied by other authors as we see below and will lead to the preferred approach to direct multi-step estimation which no longer aims to estimate the 1S model via multi-step techniques.

4. Parametric DMS Estimation Under Misspecification

The first contributions to the parametric approach to multi-step forecasting suggest some forms of model misspecification which could provide a sufficient rationale for the use of DMS. By parametric it is meant that the one-step-ahead parameters are the object of interest but that they are estimated by minimizing the functions of the multi-step errors.

Stoica and Nehorai (1989) extended the concept of direct multi-step estimation which was suggested by Findley (1983) for ARMA models:

$$A\,(L)\,y_t = C\,(L)\,\epsilon_t$$

where $A(L) = \sum_{i=0}^{p} a_i L^i$ and $C(L) = \sum_{i=0}^{p} c_i L^i$. The forecast $\widehat{y}_{T+h,h}$ is computed as the conditional expectation of y_{T+h} given y_T for the model with parameter $\theta = (a_0, \ldots, a_p, c_0, \ldots, c_p)$. Define $B_h(L) = \sum_{i=0}^{h-1} b_i L^i$ and $D_h(L) = \sum_{i=0}^{p} d_i L^i$, such that

$$C(L) = A(L)\,B_h(L) + L^h D_h(L)$$

and

$$y_t = \left(B_h(L) + \frac{D_h(L)}{A(L)} L^h \right) \epsilon_t$$

The h-step-ahead forecast error is thus given by

$$e_{T+h,h} = y_{T+h} - \widehat{y}_{T+h,h} = B_h\,(L)\,\epsilon_{T+h}$$

The authors defined the multi-step parameter estimator as that which minimizes a function of the in-sample squared multi-step forecast errors

$$\widehat{\theta}_h = \underset{\widetilde{\theta}_h \in \Theta}{\operatorname{argmin}}\, \mathsf{F}(V_{1,T}, \ldots, V_{h,T})$$

where $V_{k,T} = T^{-1} \sum_{t=1}^{T-k} e_{t+k,k}^2$, for $k = 1, \ldots, h$. They provided various algorithms to obtain the non-linear estimates.

Under the assumption that the DGP follows an ARMA(p, p), Stoica and Nehorai presented several results, namely that (1) the one-step estimator $\widehat{\theta}_1$ for $\mathsf{F}(u) = u$, is consistent and asymptotically efficient among the class of estimators whose covariance matrices depend only on the second-order properties of the data; and (2) that the only stationary point of $V_{1,\infty}$ is θ^*, the true parameter value. By contrast, they noted that the multi-step criterion may have several minima. The consequence is that, for there to be any gain from using multi-step estimation, the main assumptions have to be modified. Thus, if it is assumed that the true DGP is not known, it is still possible under weak conditions to show that $\widehat{\theta}_h$ converges to some value which leads asymptotically to the 'best' – in the sense of minimizing $\mathsf{F}(V_{1,\infty}, \ldots, V_{h,\infty})$ – forecasts. The use of DMS can therefore be justified in practice.

The authors conducted a Monte Carlo experiment in which they analysed the forecasts obtained for four models:

$$\begin{aligned}
&\text{ARMA}(3,3) && y_t - 0.95y_{t-1} + 0.81y_{t-2} - 0.7695y_{t-3}\\
& && \quad = \epsilon_t - 0.97\epsilon_{t-1} - 0.775\epsilon_{t-2} + 0.6732\epsilon_{t-3}\\
&\text{BLAR}(1) && y_t = 0.4y_{t-1} + \epsilon_t + 0.8y_{t-1}\epsilon_{t-1}\\
&\text{TMA}(3) && y_t = \begin{cases} \epsilon_t + 0.15\epsilon_{t-1} & \text{if } \epsilon_t < 0\\ \epsilon_t - 0.97\epsilon_{t-1} + 0.81\epsilon_{t-2} - 0.7857\epsilon_{t-3}, & \text{if } \epsilon_t \geq 0 \end{cases}\\
&\text{ARMA}(2,2) && y_t - 0.98y_{t-2} = \epsilon_t - 0.87\epsilon_{t-1} - 0.775\epsilon_{t-2}
\end{aligned}$$

They estimated the models over samples of size 200 and forecast over the horizons $h = 1, \ldots, 4$. The forecasting models were either an AR(4) or an AR(8), except for the ARMA(2, 2) experiment for which they tried either an AR(1) or an AR(6). The multi-step estimators were computed over the four horizons at once. Their results were that the first three models provide no rationale for the use of multi-step estimation, other than the fact that the forecast accuracy is essentially the same for IMS and DMS. By contrast the fourth model, forecast by an AR(1), does indeed provide a gain for DMS. It must be noted that the gain is for horizons 2 and 4. The slope estimates are 0.26 for IMS and 0.30 for DMS. The authors concluded that under-parameterization seems to benefit DMS.

4.1 *Discussion*

Although Stoica and Nehorai did not make explicit the difference between the model parameters and their powered-up multi-step counterparts, they showed the importance of the hypothesis of model misspecification as a justification for the use of DMS. Here, it is the same model which is used at all forecast horizons, but the estimation method matches the desired outcome. In their simulations, the authors found that an ARMA(2, 2) estimated by an AR(1) model can lead to more accurate forecasts when using DMS. Their conclusion relating to under-parameterization mirrored that of Bhansali (1999) (see Section 6); it is surprising that the very specific form of DGP they used did not strike them: it exhibited a root close to unity. It is thus possible that non-stationarity may appear as a feature benefitting DMS.

5. Efficiency in Matching Criteria for Estimation and Forecast Evaluation

Analysing the integrated ARMA (ARIMA) time series reported in Madrikakis (1982), Weiss and Andersen (1984) compared the forecasting properties of various estimation methods when the forecast accuracy criterion varies. In particular, they compared the one-step- and multi-step-ahead forecasts. They found that when a one-step-ahead forecast accuracy loss function is used, it is preferable to use one-step-ahead estimation (and then OLS and least absolute deviation seem similar in terms of both the MSFE and mean absolute error criteria). Similarly, when the forecast accuracy is measured by the absolute percentage trace of a matrix of the forecast errors at several horizons, the best among the four estimation methods which they use (the multi-step trace, one-step ahead OLS, one-step mean absolute error and one-step mean absolute percentage error) is the multi-step trace. They thus found some significant improvement from matching estimation and forecasting horizons.

Weiss (1991) built upon the earlier work on multi-step estimation for forecasting and derived conditions under which this technique is asymptotically 'optimal', in a sense that he defined. He built on the work by Johnston (1974), where, in model (6), he allowed for more lags of the endogenous variable. Weiss also defined the error terms as a function of the parameter $\boldsymbol{\theta}$,

$$\epsilon_t = \epsilon_t(\boldsymbol{\theta}) = \mathbf{y}_t - \mathbf{z}_t\boldsymbol{\theta}$$

where $\mathbf{x}_t$, in $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p}, \mathbf{x}_t)$, is a vector of variables strongly exogenous with respect to $\boldsymbol{\theta}$ (see Engle *et al.,* 1983). The model is not assumed to coincide with the DGP, and any of the following may be present: irrelevant regressors, omitted variables, serial correlation, misspecified functional form, etc. The author worked under fairly mild assumptions allowing for a uniform law of large numbers and a central limit theorem. The forecasts are given, as in Johnston (1974), as the conditional expectation computed by assuming that the model is well specified and, similarly, the forecast evaluation criterion is $\mathsf{E}[L(T, 1, \overline{h})]$, where the expectation is taken with respect to the true process, $\mathbf{Q} = \mathbf{I}$, and $L(\cdot, \cdot, \cdot)$ is defined in (7). $\widehat{\boldsymbol{\theta}}_{\text{DMS}}$ is defined as in (8), where the parameter space, $\boldsymbol{\Theta}$, is assumed compact. The inclusion of lags of $\mathbf{y}_t$ in $\mathbf{z}_t$ implies that $\widehat{\boldsymbol{\theta}}_{\text{DMS}}$ is not the simple weighted least squares estimator. Weiss assumed that a uniform law of large numbers will hold for

$$G_{T,\overline{h}}(\boldsymbol{\theta}) = \sum_{t=1}^{T-\overline{h}} L(t, 1, \overline{h}) \tag{10}$$

and that its limit coincides with that of the forecast evaluation criterion, denoted $\overline{G}_{T,\overline{h}}(\boldsymbol{\theta}) = \mathsf{E}[L(T, 1, \overline{h})]$. The author then proved that, given a minimizer of the continuous function $\overline{G}_{T,\overline{h}}(\boldsymbol{\theta})$ on $\boldsymbol{\Theta}$, which exists on a compact set and is denoted by $\widetilde{\boldsymbol{\theta}}$,

$$G_{T,\overline{h}}(\widehat{\boldsymbol{\theta}}_{\text{DMS}}) - \overline{G}_{T,\overline{h}}(\widetilde{\boldsymbol{\theta}}) \underset{T\to\infty}{\overset{\text{a.s.}}{\to}} 0$$

If the sequence of $\{\widetilde{\boldsymbol{\theta}}\}_{T=1}^{\infty}$ is identifiably unique,[1] then $\widehat{\boldsymbol{\theta}}_{\text{DMS}}$ is strongly consistent for $\widetilde{\boldsymbol{\theta}}$, i.e.

$$\widehat{\boldsymbol{\theta}}_{\text{DMS}} - \widetilde{\boldsymbol{\theta}} \underset{T\to\infty}{\overset{\text{a.s.}}{\to}} 0$$

and there exists a scaling matrix $K_{\overline{h}}$ such that

$$T^{1/2} K_{\overline{h}}(\widehat{\boldsymbol{\theta}}_{\text{DMS}} - \widetilde{\boldsymbol{\theta}}) \underset{T\to\infty}{\overset{L}{\to}} \mathsf{N}[\mathbf{0}_h, \mathbf{I}_h]\mathsf{N}\,(\mathbf{0}, \mathbf{I})$$

Thus the multi-step estimator is asymptotically optimal, in the sense that it minimizes the desired criterion function. In small samples, two opposite effects are present: the variance of the multi-step estimator should be larger than that of the one-step ahead, but the bias should be smaller. A Monte Carlo simulation thus attempted to exemplify the results for a sample of 100 observations with random initial observations. The DGP is univariate autoregressive with distributed lags (ADL),

$$y_t = \alpha_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_0 z_t + \gamma_1 z_{t-1} + \epsilon_t$$

where z_t follows a stationary AR(1): $z_t = \psi z_{t-1} + u_t$, with $|\psi| < 1$. The errors – $\{\epsilon_t\}$ and $\{u_t\}$ – are i.i.d. standard normal and independent. The intercept α_0 is set to zero and estimated in all cases. The cases studied were either a well-specified ADL(2, 1) model or one with some degree of misspecification: omitted regressors y_{t-2}, z_{t-1}, or $\{z_t, z_{t-1}\}$; or wrong functional form (estimation of the log of the data for which the intercept is non-zero). The only cases that provided a strong rationale for using the multi-step estimator are those either when $\beta_1 = 1$ and β_2 is close to zero (and not estimated) or when z_{t-1} is omitted as a regressor. Thus it seems that DMS performs better when the DGP is better modelled as a random walk than as a stationary AR(2) and when some residual autocorrelation is omitted.

Weiss (1996) analysed the forecasting properties of models which are estimated using the cost function also used for the appraisal of the forecast. The main idea is that when this criterion is quadratic, then the optimal forecast is the expectation, conditional on the information set. But this result is only valid as long as this loss function is also used for the evaluation of the forecast. Granger (1969) had considered predicting several steps into the future and recommended some techniques. For instance, letting $C_F(\cdot)$ denote the forecast evaluation cost function, if it is desired to predict y_{t+h} given $\{y_i\}_{i\leq t}$, the forecaster could minimize the in-sample bias term from a linear predictor:

$$\min_{a,b_j} \sum_t C_F\left(y_{t+h} - \sum_{j=0}^{m} b_j y_{t-j} - a\right)$$

Alternatively, if minimization is difficult to carry out, it would be sensible first to estimate $(b_1, \ldots, b_m)$, by OLS, forming $\widetilde{y}_{t+h} = y_{t+h} - \sum_{j=0}^{m} \widehat{b}_j y_{t-j}$, and then to minimize $\sum_t C_F(\widetilde{y}_{t+h} - a)$ with respect to a. Such methods were proposed because Granger thought that it should be asymptotically sensible to use the same criteria for both estimation and evaluation. In his paper, Weiss derived the optimal predictors; yet the Monte Carlo that he provided did not show substantial improvement.

5.1 *Discussion*

One of the important contributions of Weiss (1991) is that his definition of optimality is not that the estimator should have the lowest possible asymptotic variance but that it achieves the lowest in-sample (multi-step) MSFE. This shift of focus is crucial to the appraisal of DMS methods and it seems natural to evaluate a method by assessing how well it achieves the aim for which it is designed. His simulations point to the fact that, for DMS to perform better than IMS, the series must be non-stationary, either of stochastic (via unit roots) or of deterministic (since location shifts imply residual autocorrelation) form.

6. Design of Non-parametric Forecasting Models

This first strain of direct estimation focuses on fitting different models for forecasting at different horizons. Research along these lines attempts to establish reasonable 'good' criteria for choosing the order p of the 'best' AR(p) to use for forecasting. The 'non-parametric' terminology is explained in Bhansali (1999).

Findley (1983) provided a theoretical rationale for adapting the forecasting models to the forecasting horizon and suggested one type of technique which he applied to some standard time series from Box and Jenkins (1976). The author started by considering the case when an AR(1) model is used for the prediction of a variable h steps ahead. Denoting by $\{\rho_k\}$ the autocorrelation sequence of the process $\{y_t\}$, the parameter ψ_h which minimizes the MSFE

$$\mathsf{E}\left[(y_{T+h} - \widehat{y}_{T+h,h})^2\right] \tag{11}$$

where $\widehat{y}_{T+h,h} = \psi_h y_T$, is simply the autocorrelation $\psi_h = \rho_h$, in the stationary case. If $\{y_t\}$ does indeed follow an AR(1), $y_t = \phi y_{t-1} + \epsilon_t$, where $\phi < 1$, then, naturally, we need to choose $\psi_h = \phi^h$ and $\phi = \rho_1$. If $\{y_t\}$ follows any other process but we still fit an AR(1) as above, in order to minimize (11), we must set

$$\begin{aligned} \phi &= (\rho_h)^{1/h} && \text{if } h \text{ is odd or } \rho_h > 0 \\ \phi &= 0 && \text{if } h \text{ is even and } \rho_h < 0 \end{aligned}$$

Thus the 'optimal' model depends on the desired lead in forecasting. Notice that if h is even and $\rho_h < 0$, it is preferable not to fit an AR(1) model but rather to use $\psi_h = \rho_h$. It therefore seems that the formula most appropriate to multi-step forecasting cannot always be derived from an ARMA model. The results would asymptotically be the same if the estimators were computed to maximize the forecast log-likelihood. Findley remarked that when the forecast accuracy criterion combined several horizons, the degree of complexity was much higher. In order to improve forecast accuracy, it may seem desirable to use several lags of the variable. Findley suggested an h-step Akaike information criterion (AIC$_h$) in order to select the order of the AR(p) to be fitted (for p smaller than some $p_{\max}$). The order p is thus given by

$$p = \underset{1 \leq p \leq p_{\max}}{\operatorname{argmin}} \{\text{AIC}_h(p)\}$$

where

$$\mathrm{AIC}_h(p) = T_0 \log\left[2\pi SSQ\left(\widehat{\phi}_1, \ldots, \widehat{\phi}_p\right)/T_0\right] + T_0 + 2\,(p+1)$$
$$T_0 = T - p_{\max} - h + 1$$

and $(\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$ is computed as the set of coefficients which minimizes the in-sample sum of the squared h-step-ahead residuals

$$SSQ\left(\phi_1, \ldots, \phi_p\right) = \sum_{t=p_{\max}}^{T-h} \left(y_{t+h} - \sum_{k=1}^{p} \phi_k y_{t-k+1}\right)^2$$

The author applied his results to two standard time series: series C and E from Box and Jenkins (1976), fitting autoregressive models using the AIC$_h$ criterion. The results exhibit an average gain for the proposed multi-step methods in terms of MSFE of about 4% for series C at horizons 5 and 10 and, respectively, 2.6% and 10.6% for series E at horizons 5 and 10.

Liu (1996) suggested modification of the standard fitting criteria for the order of an autoregressive process to allow for the inclusion of multi-step forecast errors. He proposed partitioning the data set into non-overlapping vectors of length h, where h is the maximum desired forecast horizon. Estimating the resulting vector autoregression (VAR) by weighted least squares was shown asymptotically to lead to the same estimates as those of a univariate model, but with a loss of efficiency. In a Monte Carlo simulation, for samples of size 80 and 240, Liu compared the ratios of 2- and 4-step-ahead root MSFEs. The results show little improvement when using the multi-step methods, whether the data were generated by either a zero-mean stationary AR(1) or an ARI(1, 1). The author applied his methods to forecasting the quarterly US (174 observations) and monthly Taiwan (192 observations) unemployment rates, the log of quarterly real US GNP (179 observations) and the monthly US consumer price index for food (241 observations). Several overlapping samples for each data set were used where estimation was carried out for fixed sample sizes of, respectively, 100, 120, 100, 160 observations. Three main results emerge: first, when the one-step method is preferred, the loss from using a DMS method is low, except when the multi-step order is determined by a modified Fisher information criterion for which the loss can be up to 12.5%; second the multivariate procedure is always preferred for the trending variables in levels (the US GNP and food index), but not necessarily in differences; and third the DMS method is preferred for the monthly Taiwan unemployment rate but not for the quarterly US rate. For the latter result, the author suggested as an explanation that monthly data exhibited more time dependence.

Bhansali (1999) surveyed the developments in multi-step criteria for the design of forecasting models. He first distinguished two different approaches: a parametric and a non-parametric. In the former, the modeller attempts to establish what the true DGP is, and estimates its k parameters via some multi-step technique (for instance, maximum likelihood); by contrast a non-parametric procedure approximates the unknown DGP by some process whose number of parameters is allowed to diverge, say $k(T)$, where T is the sample size and $k(T) = \mathrm{o}(T)$. Assume that a process $\{y_t\}$ is approximated or modelled as a linear function of k lags, so that at an end-of-sample forecast origin T, it is desired to predict y_{T+h}, where $h \geq 1$. The resulting forecast $\widetilde{y}_{T+h,h}$ is written as

$$\widetilde{y}_{T+h,h} = \sum_{i=0}^{k} \widetilde{\alpha}_{h,i} y_{T-i}$$

where the $\widetilde{\alpha}_{h,i}$ are estimated by regressing y_{t+h} on $(y_t, \ldots, y_{t-k})$ from a hypothesized model (the forecast-generating process), for fixed $h \geq 1$,

$$y_{T+h} = \sum_{i=0}^{k} \alpha_{h,i} y_{T-i}$$

For notational simplicity, the dependence of the $\alpha_{h,i}$ on k is omitted. Define the unconditional MSFE

$$\widetilde{V}_{h,k}^{\text{DMS}} = \mathsf{E}\big[(y_{T+h} - \widetilde{y}_{T+h,h})^2\big] \tag{12}$$

Similarly, letting $y_{T+1,1} = \sum_{i=0}^{k} \alpha_{1,i} y_{T-i}$ and noting that

$$\begin{aligned} y_{T+2} &= \alpha_{1,0} y_{T+1,1} + \alpha_{1,1} y_T + \alpha_{1,2} y_{T-1} + \ldots \\ &= \sum_{i=0}^{k-1} (\alpha_{1,0}\alpha_{1,i} + \alpha_{1,i+1}) y_{T-i} + \alpha_{1,0}\alpha_{1,k} y_{T-k} \end{aligned}$$

it is possible, by iterated substitution, to find a set of non-linear functions $\beta_{h,i}$ of the set $\{\alpha_{1,i}\}$ such that

$$y_{T+h} = \sum_{i=0}^{k} \beta_{h,i} y_{T-i}$$

Denote by $\widehat{\beta}_{h,i}$ the function of the estimated $\widetilde{\alpha}_{1,i}$ so that $\widehat{y}_{T+h,h} = \sum_{i=0}^{k} \widehat{\beta}_{h,i} y_{T-i}$. Then, the IMS MSFE is defined as

$$\widehat{V}_{h,k}^{\text{IMS}} = \mathsf{E}\Big[\big(y_{T+h} - \widehat{y}_{T+h,h}\big)^2\Big]$$

Note that $\{y_t\}$ can expressed as an autoregressive process of order p according to the Wiener–Kolmogorov theorem, where p can be infinite. Then, if $k \geq p$, the theoretical (for known parameters) MSFEs coincide for DMS ($V_{h,k}^{\text{DMS}}$) and IMS ($V_{h,k}^{\text{IMS}}$); but if $k < p$, the latter is larger than the former, which in turn is larger than the 'true' MSFE from the correct – potentially infinitely parameterized – model (see Bhansali, 1996). Define γ_i as the ith autocorrelation of $\{y_t\}$ for $i \geq 1$ and γ_0 as its variance (in stationary processes); then using an AR(1) as a forecasting model,

$$\frac{V_{2,1}^{\text{IMS}}}{V_{2,1}^{\text{DMS}}} = 1 + \frac{\left[\gamma_2 - (\gamma_1)^2\right]^2}{1 - (\gamma_2)^2} \geq 1$$

where the equality arises when the model is well specified. Similarly, it can be shown that if the process follows an MA(1) with parameter θ, $V_{2,1}^{\text{IMS}}/V_{2,1}^{\text{DMS}} = 1 + [\theta/(1+\theta^2)]^4 > 1$.

Hence, for a well-specified model, the 1S estimation procedure is asymptotically equivalent to maximum likelihood and, in the case of Gaussian processes, achieves the Cramér–Rao bound; yet this is not the case when $k \neq p$. By contrast, DMS is asymptotically inefficient for a well-specified model. However, if $k(T) \to \infty$, the distributions of the DMS and IMS estimators coincide for $T \to \infty$, under some regularity conditions (see Bhansali (1993) when $k(T) = \mathrm{o}(T^{1/3})$). In analysing the ARMA(1, 1) model

$$y_t - \rho y_{t-1} = \epsilon_t - \theta \epsilon_{t-1}$$

Bhansali noted that the two-step-ahead forecast is given by

$$y_{t+2} = -\rho\,(\theta - \rho) \sum_{i=1}^{\infty} \theta^i y_{t-i} = \tau\,(1 - \theta L)^{-1}\, y_t \tag{13}$$

so that he recommended minimization of the in-sample sum of squared forecast errors,

$$\sum [y_{t+2} - \tau\,(1 - \theta L)^{-1}\, y_t]^2$$

for $(\tau,\ \theta)$ rather than for the original parameters $(\rho,\ \theta)$ since it is the multi-step parameters which present an interest. Another justification was given by Stoica and Soderstrom (1984) who showed that the parameter estimates $(\widehat{\tau}, \widehat{\theta})$ are unique whereas $(\widehat{\rho}, \widehat{\theta})$ may not be so. We therefore call the model (13), with parameters $(\tau,\ \theta)$, the forecast-generating process (FGP).

When the process to forecast or the model used is non-stationary – like the structural time series in Harvey (1993) – Haywood and Tunnicliffe-Wilson (1997) extended the work by Tiao and Xu (1993) to direct multi-step estimation of spectral densities. This is also the focus of the paper by Hurvich (2002): he analysed multi-step forecasting using the fractional exponential FGP which was estimated in the frequency domain. In this paper, the number of regressors is obtained by an information criterion based on the multi-step forecast error but the author provided no forecast accuracy results. Bhansali reviewed the different criteria which could be used for deciding on the lag length k to be used for forecasting and noted that some asymptotic efficiency could be shown for the MSFE obtained by DMS when k was treated as a random variable function of a modified AIC. Finally, Bhansali concluded that there exists a rationale for DMS when the model is under-parameterized for the DGP or when the latter is complex or belongs to a class admitting an infinite number of parameters. He remarked also that even if a model is fitted to the data and seems to pass the traditional diagnostic tests, there might be a DMS FGP which, because it explicitly allows for moving average errors, improves and robustifies the forecasting performances.

Schorfheide (2005) extended and confirmed these results by presenting the case of local misspecification, whereby the disturbances exhibited serial correlation that asymptotically vanished.

Bhansali (2002) applied DMS estimation to a Monte Carlo experiment of seasonally adjusted autoregressive AR(9) time series estimated over a sample of 99 observations. The FGP was selected using the criterion in Shibata (1980), but this often led to selecting a model of order 0. The author concluded that his simulation did not seem to advocate the use of direct multi-step estimation and assumed that removing the seasonality may have damped the serial dependence of the process, or that the sample used was too small, or finally that this result may have simply depended on the specific series used in the simulation.

More recently Ing (2003) has analysed the estimation of stationary AR(p) processes via a misspecified AR(k) model, and, contrary to Bhansali (1996), did not assume independence between the estimation sample and the realizations to forecast. For $k \geq p$, IMS is then asymptotically more efficient than DMS (in terms of MSFE) and for both methods a lower k is more efficient as long as it is not lower than p. By contrast, Ing showed that when $k < p$, for given h and k,

$$\lim_{T \to \infty} (\text{MSFE}_{\text{IMS}} - \text{MSFE}_{\text{DMS}}) > 0$$

Non-parametric direct multi-step estimation was also the focus of Clements and Hendry (1996) and Chevillon (2005). But these authors analysed estimation for forecasting rather than the design of the DMS FGP. They shed light on the dynamic properties leading direct estimation to improve accuracy and hence we present their contributions in Section 8 where we review the features advocating the use of DMS. In another paper in 2004, an extension of the earlier work was presented and Ing provided an algorithm for the design of an FGP which proved asymptotically optimal in the sense of minimizing the multi-step forecast error criterion presented in (10),

$$G_{T,h}(\boldsymbol{\theta}_k) = \sum_{t=m_h}^{T-h} L(t, 1, h) = \sum_{t=m_h}^{T-h} (y_{t+h} - \widehat{y}_{t+h,h})^2$$

which the author related to a multi-step generalization of Rissanen's accumulated prediction errors (APE) (see Rissanen, 1986). Contrary to (10), Ing allowed for $m_h > 1$ to make sure that the forecasting models are defined. Here, $\widehat{y}_{t+h,h}$ is generated using an AR(k) with a vector of parameters $\boldsymbol{\theta}_k$ which differs depending on whether IMS or DMS estimation and forecasting are carried out. APE criteria differ from standard least squares in so far as $\widehat{y}_{t+h,h}$ is computed from a model estimated solely on the basis of data available up to t: this focuses on in-sample accuracy of recursive *ex ante* forecasting. Hence APE is similar to some multi-step recursive least squares.

In practice, the suggested algorithm aims at choosing an order k for the FGP and a method (iterated or direct). The three recommended steps are as follows: (1) find a minimum $k_{\min}$ for the autoregressive-order k^{IMS} by minimizing $G_{T,1}(\boldsymbol{\theta}_k)$ where the IMS predictor is used and then (2) obtain the order k^{DMS} by minimizing $G_{T,h}(\boldsymbol{\theta}_k)$ with direct prediction for $1 \leq k \leq K$ for some K, and similarly k^{IMS} using the criterion with IMS forecasting and $k_{\min} \leq k \leq K$; finally (3) the recommended technique and FGP are the combination that leads to the lower $G_{T,h}(\boldsymbol{\theta}_k)$. Ing therefore compared, for a given horizon, various models within a given class (stationary and autoregressive) and chose amongst them using an in-sample forecasting criterion.

In this paper, Ing assumed that the DGP is a stationary AR(p) of finite order with innovations whose moments exist at least up to some order greater than 8 and denoted by p_h the autoregressive order of the linear predictor of y_{t+h} based on an infinite past and known parameters. Under these hypotheses, it was shown that for a given h, with the same notation as above for the unconditional MSFE $V_{h,k}^{\text{IMS}}$ and with $\sigma_h^2 = \mathsf{E}[(y_t - y_{t-h})^2]$,

$$\begin{aligned} T\left(V_{h,k}^{\text{IMS}} - \sigma_h^2\right) &= f_h^{\text{IMS}}(k) + \mathrm{O}(T^{-1/2}) \quad \text{for } k \geq p_1 \\ T\left(V_{h,k}^{\text{DMS}} - \sigma_h^2\right) &= f_h^{\text{DMS}}(k) + \mathrm{O}(T^{-1/2}) \quad \text{for } k \geq p_h \end{aligned}$$

Several asymptotic properties are discussed regarding the monotonicity of f_h^{IMS} and f_h^{DMS} for $k \geq p$ which is valid except on a non-empty subset. The main results concern the APE criterion given by

$$G_{T,h}\left(\widehat{\boldsymbol{\theta}}_k^{\text{IMS}}\right) - \sum_{t=m_h}^{T-h} (y_{t+h} - y_t)^2 \underset{\text{a.s.}}{=} \sigma^2 f_h^{\text{IMS}}(k) \log T + \mathrm{o}(\log T) \quad \text{for } k \geq p_1 \tag{14}$$

and the corresponding property holds for DMS and $k \geq p_h$. An actual expression for (14), when scaled by T^{-1}, is provided up to o(1), which allows the author to state that for $K \geq p_1$ the suggested algorithm yields the asymptotically efficient forecast in the sense of minimizing the loss criterion ($j \in$ {IMS, DMS} and $p_{(\text{IMS})} = p_1,\ p_{(\text{DMS})} = p_h$)

$$L(h) = \begin{cases} \lim\limits_{T\to\infty} T\left(V_{h,k}^{(j)} - \sigma_h^2\right) & \text{for } p_{(j)} \leq k \leq K \\ \infty & \text{if } k < p_{(j)} \end{cases}$$

Thus, an infinite loss is attributed to an underparameterized FGP and in-sample minimization of $G_{T,h}(\boldsymbol{\theta}_k)$ and is then asymptotically valid as a criterion for multi-step forecast accuracy. The author then proves that this is also the case when intermediate lagged values (between 1 and k) are allowed to be excluded from the regressor set for either of the IMS and DMS models.

6.1 *Discussion*

There is an extended literature on non-parametric DMS where the authors focus particularly on choosing a forecast-generating model by designing criteria or estimating long-memory time series. Results concur to show that these methods need reasonably large samples and strong time dependence – hence a recent focus of researchers on forecasting fractionally integrated time series. The multivariate framework in Liu (1996) has the disadvantage that it partitions the set of observations in non-overlapping subsets and thus loses a lot of information. It therefore cannot be used when only one forecast horizon matters, and it is not clear that estimating all horizons at once will yield any better results than forecasting each separately, thus taking full advantage of the DMS framework. The definition of non-parametric DMS by Bhansali is constrained to model design or choice. He thus omits work on the *estimation* approach to DMS where the focus is not, as in the parametric approach, on the 1S parameters but on the multi-step parameters that matter for forecasting. In this framework and for stationary processes, Ing (2003, 2004) justified some of the previous observations and suggested a procedure that led to choosing the optimal model (in the sense of non-underparameterized model which minimized the asymptotic *unconditional* MSFE) by in-sample minimizing of the recursive multi-step forecast errors. DMS is then efficient for underparameterized models. Unfortunately, although the asymptotic variances of the multi-step forecast errors are derived, this work does not seem to establish what specific features of the DGP render iterated or direct methods more attractive (this will constitute the purpose of Section 8) and stationarity has to be assumed throughout, in particular posterior to the forecast origin.

7. Robustness of Multi-step Forecasts From ARMA Models

Tiao and Xu (1993) developed an extensive analysis of the properties of the DMS forecasts generated by an exponential smoothing formula – the FGP – which was estimated when the (true) DGP followed an ARIMA(p, $d = 0$ or 1, q). They thus extended the results by Cox (1961) and showed that multi-step estimation may be preferable. They motivated their study by comparing IMS and DMS forecasting properties for series A, from Box and Jenkins (1976): they fit an ARIMA(0, 1, 1) model where the moving average parameter, θ, was estimated by minimizing the in-sample multi-step squared residuals implied by the exponential smoothing formula. The estimates hence depended on the forecast horizon. The authors used the mean corrected series and let the sample size vary from 101 to 157 observations. They reported the ratios of the average (over the resulting 57 outcomes) squared forecast error for the IMS over those from the DMS estimation technique. The forecast horizon varied from 2 to 40 and the ratio first decreased with the lead (thus benefitting IMS) until horizon $h = 7$, and then it established itself between about 1.3 and 1.6. It must be noted, though, that the estimate $\widehat{\theta}_h$ increased with h and, from observation $h = 15$ onwards, it was unity, thus implying that the forecast was simply the sample average.

The authors extended the framework in Cox (1961) to a process $\{y_t\}$ which follows an ARIMA(p, d, q),

$$\phi(L)(1-L)^d y_t = \xi(L)\epsilon_t \tag{15}$$

where $\phi L)$ and $\xi(L)$ are polynomials – whose roots are stationary – of orders, respectively, p and q, and d is either 0 or 1. The aim is to analyse the robustness of the h-step-ahead forecasts when these are obtained by the exponential smoothing formula

$$\widehat{y}_{T+h,h} = (1-\theta_h)\sum_{t=0}^{T-1}\theta_h^t y_{T-t} \tag{16}$$

and the forecast error is given by

$$\widehat{e}_{T+h,h} = y_{T+h} - \widehat{y}_{T+h,h} \tag{17}$$

The asymptotic h-step-ahead MSFE is, in $\mathbb{R}\cup\{-\infty, +\infty\}$,

$$\sigma^2(h,\theta) = \lim_{T\to\infty} \mathsf{E}\left[\widehat{e}^2_{T+h,h}\right]$$

The authors showed that the MSFE can be decomposed into the sum of the variance of the h-step-ahead forecast errors under the 'true' model plus the squared bias introduced by the misspecification. This allowed them to derive the exact formula for $\sigma^2(h, \theta)$, which exists for $\theta \in (-1, 1)$, when $d = 1$, and for $\theta \in (-1, 1]$, for $d = 0$.

If $\{y_t\}$ follows an ARIMA(1, 0, 1), then

$$y_t - \phi y_{t-1} = \epsilon_t - \xi \epsilon_{t-1}$$

which is referred to as the (ϕ, ξ) model and the forecasting model is then denoted by $(1, \theta)$. Tiao and Xu, then, derived the minimum of $\sigma^2(h, \theta)$, for given h, and it is given by

$$\theta_h^* = \begin{cases} (1 - \sqrt{c})(\phi + c)^{-1} & \text{for } (\phi, \xi) \in \mathbf{S} \\ 1 & \text{otherwise} \end{cases}$$

where $c = (1 + \xi)(\phi - \xi)(1 - \phi\xi)[(1 + \phi)\phi^{h-1} - 1]$ and $\mathbf{S}$ is some region of $[-1, 1] \times [-1, 1]$ which they defined. Let $r(h; \phi, \xi)$ be the ratio of $\sigma^2(h, \theta_h^*)$ over the MSFE implied by the true model; it is a measure of efficiency loss. The authors showed that $r(1; \phi, \xi) < 1.2$ over a wide region around $\phi = \xi$, or when $\phi > \xi > 0$, or when $\frac{2}{3} < \phi \leq 1$; and it is moderate over a large part of the parameter space, as is often the case in empirical work, and which is one of the reasons for the widespread use of the exponential smoothing formula. When (ϕ, ξ) vary, θ_h^* is unity when ϕ is negative, or when $\xi > \phi$. As regards the behaviour with respect to h, the supremum of $r(h; \phi, \xi)$, when $\phi > \xi > 0$, increases with the horizon but is bounded as $h \to \infty$ by 4/3. When comparing DMS and IMS forecasting performances, the authors mentioned that the ratio $\sigma^2(h, \theta_1^*)/\sigma^2(h, \theta_h^*)$ is increasing in h for $\phi > \xi > 0$ and it tends to 2 as the horizon goes to infinity.

Under the general ARIMA case, in (15), Tiao and Xu then proved the consistency of the estimate $\widehat{\theta}_h(T)$ of θ_h^* obtained by minimizing $(T - h)^{-1} \sum_{t=1}^{T-h} \widehat{e}_{t+h,h}^2$, the in-sample average of the squared forecast errors; they showed that, under some regularity assumptions,

$$\widehat{\theta}_h(T) \underset{T \to \infty}{\to} \theta_h^*$$

where θ_h^* is a – the, if unique – minimizer of $\sigma^2(h, \theta)$ over $(-1, 1]$. This result extends to forecasts generated from the FGP

$$(1 - L)^{b_1}(1 - L^s)^{b_2} y_t = (1 - \theta_1 L)(1 - \theta_2 L^s) u_t$$

where $\{u_t\}$ is assumed to be i.i.d. Gaussian white noise, $s \geq 1$, $b_1 = 0$ or 1, $b_2 =$ 0 or 1, $b_1 + b_2 > 0$, and the DGP of the series is

$$\phi(L)(1 - L)^{d_1}(1 - L^s)^{d_2} y_t = \xi(L)\epsilon_t$$

The FGP includes here, *inter alia*, the ARIMA(0, 2, 2) non-stationary smooth trend model ($s = 1$, $b_1 = b_2 = 1$), and the multiplicative non-stationary seasonal models ($s = 12$, $b_1 = b_2 = 1$ and $s = 12$, $b_1 = 0$, $b_2 = 1$).

Focusing on the dependence relation between the parameter estimates for varying forecast horizons, the authors showed that if the true DGP is $(1-L)y_t = (1-\theta_0 L)\epsilon_t$, where $\epsilon_t \sim \mathsf{IN}(0, \sigma_\epsilon^2)$, then

$$T^{1/2}\left(\frac{[\widehat{\theta}_1(T)-\theta_0]}{[1-\widehat{\theta}_1(T)]^{1/2}}, \frac{[\widehat{\theta}_2(T)-\widehat{\theta}_1(T)]}{1-\widehat{\theta}_1(T)}, \ldots, \frac{[\widehat{\theta}_h(T)-\widehat{\theta}_{h-1}(T)]}{1-\widehat{\theta}_1(T)}\right)' \underset{T\to\infty}{\overset{L}{\to}} \mathsf{N}[\mathbf{0}_h, \mathbf{I}_h] \tag{18}$$

This means that when the FGP and DGP coincide, the loss of efficiency from using a multi-step estimation procedure is

$$\frac{\mathrm{Var}\left[T^{1/2}(\widehat{\theta}_h(T)-\theta_0)\right]}{\mathrm{Var}\left[T^{1/2}(\widehat{\theta}_1(T)-\theta_0)\right]} = 1 + (h-1)\left(1-\theta_0^2\right) \quad \text{for } h \geq 1$$

The results from (18) imply that multi-step estimation can be used to generate diagnostic tests. The authors suggested two of them and compared them to the Box–Ljung and Dickey–Fuller statistics. Yet, although the results seemed promising in small samples, they were not decisive.

The contribution of Tiao and Xu is thus to show that direct multi-step estimation can lead to more efficient forecasts when the model is misspecified. Yet, when the forecasting model and the DGP coincide, it is still asymptotically preferable to use IMS in large samples since DMS leads to an efficiency loss.

Tiao and Tsay (1994) provided some theoretical and empirical considerations for the use of multi-step ('adaptive') estimation for forecasting. Their focus was on long-memory processes which could be represented by fractionally integrated ARMA (ARFIMA) models. They compared the resulting forecasts to those obtained via single-step or multi-step estimation of a stationary ARIMA model,

$$(1-\alpha L)\, y_t = (1-\rho L)\, \epsilon_t$$

where $|\alpha| < 1$, $|\rho| < 1$ and ϵ_t is not modelled since it is known that the FGP is misspecified for the DGP. The resulting h-step-ahead forecasts and forecast errors are given by

$$\widehat{y}_{T+h,h} = \begin{cases} \alpha y_T - \rho \epsilon_T & \text{for } h = 1 \\ \alpha^{h-1}\widehat{y}_{T+1,1} & \text{for } h \geq 2 \end{cases}$$

and

$$\widehat{e}_{T+h,h} = y_{T+h} - \alpha^{h-1}(\alpha y_T - \rho\epsilon_T)$$

which imply that the variances of the forecast errors are

$$\mathrm{Var}[\widehat{e}_{T+h,h}] = \begin{cases} \sigma_\epsilon^2 & \text{for } h = 1 \\ \sigma_y^2\left(1-2\alpha^h\gamma_h+\alpha^{2h}\right)+\alpha^2(h-1)\rho^2\sigma_\epsilon^2 & \\ \quad +2\alpha^{h-1}\rho\mathrm{Cov}\left[y_{T+h}-\alpha^h y_T, \epsilon_t\right] & \text{for } h \geq 2 \end{cases} \tag{19}$$

where σ_y^2 and σ_ϵ^2 are the variances of y_t and ϵ_t and γ_h is the lag h auto-correlation of y_t. Thus, multi-step estimation would lead to minimizing the variance in (19), and optimal values would depend on the horizon. Tiao and Tsay compared the Monte Carlo MSFEs from one-step and multi-step estimation to the 'true' forecast error variances obtained using the DGP

$$(1-L)^d y_t = u_t \text{ and } u_t \sim \mathsf{IN}\left(0, \sigma_u^2\right)$$

for the two values $d = 0.25$ and 0.45 (i.e. close to the non-stationarity coefficient of 0.5). They did not mention the sample size used for estimation, but reported the forecast statistics up to 200 steps ahead. Their results showed that the loss from using the misspecified DMS ARIMA is never more than 5% in terms of MSFE and, in fact, almost always less than 1% when $d = 0.25$. The gain from DMS versus IMS is not significant – yet positive – when $d = 0.25$, but it is so, and rapidly increasing with the horizon, for almost all non-stationary processes: it is about 6% for $h = 4$, 13% at $h = 10$, 26% at $h = 20$, 57% at $h = 50$ and 70% for $h = 100$ or 200. In practice, though, the distribution of $\{y_t\}$ is not known and (19) cannot be computed; yet the modeller can still compute some estimates by minimizing the in-sample squares of the forecast errors $\widehat{e}_{t+h,h}$ for $t = 1, \ldots, T$.

Tiao and Tsay then applied their method to the prediction of the differences in the series of the US monthly consumer price index for food from January 1947 to July 1978, which have been reported in previous studies to be well modelled by an ARFIMA(0, 0.423, 0) process. Using samples of 80 observations, the authors estimated the models used for the Monte Carlo and compared the resulting empirical MSFEs, computed as the average of the squared out-of-sample forecast errors. Their results strongly favoured multi-step estimation of an ARIMA(1, 1) over the other two techniques at all horizons and especially at large ones ($h \geq 40$). Tiao and Tsay concluded by noting that one of the advantages of DMS is its estimation simplicity and the fact that it can be extended to forecast linear aggregates of future observations.

In his comment on Tiao and Tsay (1994), Peña (1994) suggested another case when multi-step estimation leads to better forecasting properties. He assumed that the DGP presents an additive outlier (unknown to the modeller),

$$x_t = z_t + \omega.1_{\{t=T\}}$$

and that Δz_t follows an AR(1) process without intercept and defines

$$y_t = \Delta x_t = \Delta z_t + \omega\left(1_{\{t=T\}} - 1_{\{t=T+1\}}\right)$$

Assume that the autoregressive coefficient of $\{y_t\}$, α, is estimated by minimizing the in-sample multi-step forecast errors. Denote the resulting estimator by $\widehat{\alpha}_h$, where

$$\widehat{\alpha}_h = \left(\frac{\sum y_{t+h} y_t}{\sum y_t^2}\right)^{1/h} = r_h^{1/h}$$

Therefore, in the presence of the outlier

$$\widehat{\alpha}_h = \left(\frac{\omega\left(z_{T+h} + z_{T-h} - z_{T+h+1} - z_{T-h+1}\right) + \sum z_{t+h} z_t}{2\omega^2 + 2\omega\left(z_T - z_{T+1}\right) + \sum z_t^2}\right)^{1/h} \quad \text{for } h > 1$$

$$\widehat{\alpha}_1 = \left(\frac{\omega\left(z_{T+1} + z_{T-1} - z_{T+2} - z_T\right) + \sum z_{t+1} z_t}{2\omega^2 + 2\omega\left(z_T - z_{T+1}\right) + \sum z_t^2}\right)$$

and $\widehat{\alpha}_1$ is hence more affected by the outlier than $\widehat{\alpha}_h (h > 1)$. Multi-step estimation may thus provide more robust estimates of the true parameters. Peña noted also that such an estimation method can be used for diagnostic purposes. Breaks are also the focus of Chevillon (2006) who analysed the context of unnoticed *location shifts* (i.e. structural breaks affecting the mean of the process) occurring a few periods prior to the forecast origin. He showed that DMS then acts as an intermediate strategy between congruent modelling and intercept correction by taking better account of the dynamic properties of the data.

In a comment about the computation of forecast intervals, Tsay (1993) suggested the use of multi-step (adaptive, for him) forecasting. His rationale was that all statistical models are imperfect representations of the reality and that, when it comes to forecasting, local approximations are more relevant than global ones. The two main implications of these remarks are that the maximum likelihood principle does not apply and that, since 'different forecast horizons have different local characteristics', different models should be fitted for each forecast. The author then considered forecasting the US quarterly unemployment rate as in Chatfield (1993). The estimates were computed by minimizing the in-sample sum of squares of the multi-step residuals obtained by assuming that the DGP could be approximated by an AR(2) model. This method results in non-linear estimation. This follows the technique used in Tiao and Tsay. Tsay provided the point forecasts up to 12 steps ahead together with the 95% prediction interval from the in-sample empirical distribution of the multi-step residuals. He compared his results with those obtained by fitting an AR(1) and an ARIMA(1, 1, 0) model. Estimation over a sample of 48 observations led to the AR(2)

$$y_t = 0.4409 + 1.5963 y_{t-1} - 0.6689 y_{t-2} + \epsilon_t$$

which implies that the series is nearly integrated with a root of 0.9274. The author found that the multi-step point forecasts were more accurate than those obtained by the other models. The prediction interval did not necessarily increase with the horizon for the 'adaptive' forecast and it had the same amplitude as that of the AR(2) but, contrary to the latter, always (except at 12 steps ahead) contained the true outcome. The forecast interval of the ARIMA model also contains the realized value but it is much larger than that of the previous two models. The author concluded that multi-step estimation did indeed yield a positive outcome.

In their paper, Lin and Tsay (1996) studied whether using cointegrating properties improves long-term forecasting. In an empirical analysis, they compared several forecasting techniques to an 'adaptive' procedure of multi-step forecasts, which they used as a benchmark since it did not postulate the existence of a true model. They used a non-stationary VAR(p) model for the vector of n variables,

$$\mathbf{x}_t = \boldsymbol{\tau} + \sum_{i=1}^{p} \boldsymbol{\Upsilon}_i \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t$$

where $\epsilon_t \sim \mathsf{IN}(\mathbf{0}, \Sigma)$. It is assumed that, letting $\boldsymbol{\Upsilon}(L) = \sum_{i=1}^{p} \boldsymbol{\Upsilon}_i L^i$, the series $\{\boldsymbol{\Psi}_i\}_{\infty}^{0}$ is such that $[\boldsymbol{\Upsilon}(L)]^{-1} = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_i L^i$. The h-step-ahead forecast error is then

$$\mathbf{e}_{T+h,h} = \sum_{i=0}^{h-1} \Psi_i \epsilon_{T+h-i}$$

The DMS multi-step parameter estimates are obtained via minimizing the in-sample sum of the squared $\mathbf{e}_{t,h}$. This implies a non-linear function of the elements of $\{\boldsymbol{\Upsilon}_i\}_p^1$. For simplicity, Lin and Tsay suggested using the least squares projection of $\mathbf{x}_t$ onto the space spanned by $(\mathbf{x}_{t-h}, \ldots, \mathbf{x}_{t-h-p+1})$ and a constant, for $t = h + p, \ldots, T$. The computing time of this alternative estimator was much lower.

Lin and Tsay compared their DMS forecasts with those obtained by cointegrated VARs for seven financial and macroeconomic data sets. The vector processes were of a dimension varying from 3 to 5 and were estimated over samples of 230 to 440 observations. The criterion used for analysis was the square root of the average trace of the MSFE. Their results showed that multi-step techniques provide a greater forecast accuracy (up to a 60% gain), but for long horizons (beyond 50) only two of the series still exhibited a gain from using DMS. The authors found it difficult to account for these results.

7.1 *Discussion*

The papers summarized here provide a vast array of justifications for, and successful examples of the use of, DMS methods. They confirm that three main types of model misspecification benefit direct multi-step forecasting, namely misspecified unit roots, neglected residual autocorrelation and omitted location shifts – although the latter two can be thought of as representations of the same phenomenon. They also suggest that the success or failure of DMS can be used as a model specification test. Peña showed that DMS estimates were more robust to additive outliers than one-step, but the practical use of this feature for forecasting may not be so significant in practice if indeed a shift occurs. Finally, Lin and Tsay, like Bhansali, contrasted the two ways to proceed with DMS estimation and forecasting: via either (1) using the same FGP for both IMS and DMS, the DMS estimated being computed by minimizing the implied in-sample h-step residuals, which can be non-linear functions of the FGP parameters; or (2) using a different model at each horizon where it is the multi-step parameters – defined as the coefficients from a projection of y_t on the information set up to time $t - h$ – which are estimated.

8. When Does DMS Win?

Clements and Hendry (1996) developed an extended analysis of multi-step estimation for stationary and integrated processes. Their focus is on VAR(1) models as in

$$\mathbf{x}_t = \boldsymbol{\Upsilon} \boldsymbol{x}_{t-1} + \epsilon_t \qquad (20)$$

where the n-vector process $\{\boldsymbol{\epsilon}_t\}$ satisfies $\mathsf{E}[\boldsymbol{\epsilon}_t] = \mathbf{0}$. From an end-of-sample forecast origin T,

$$\mathbf{x}_{T+h} = \boldsymbol{\Upsilon}^h \mathbf{x}_T + \sum_{i=0}^{h-1} \boldsymbol{\Upsilon}^i \boldsymbol{\epsilon}_{T+h-i} \tag{21}$$

and the IMS and DMS forecasts are given, respectively, by

$$\widehat{\mathbf{x}}_{T+h} = \widehat{\boldsymbol{\Upsilon}}^h \mathbf{x}_T \quad \text{and} \quad \mathbf{x}_{T+h} = \widetilde{\boldsymbol{\Upsilon}}_h \mathbf{x}_T$$

where $\widehat{\boldsymbol{\Upsilon}}$ and $\widetilde{\boldsymbol{\Upsilon}}_h$ are the estimators of $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Upsilon}^h$ obtained by minimizing, respectively, the one-step-ahead and h-step-ahead in-sample forecast errors. The authors noted that the relative accuracy of DMS versus IMS was given by that of the powered estimate versus the estimated power. Direct estimation of $\widetilde{\boldsymbol{\Upsilon}}_h$ has therefore some potential when $\widehat{\boldsymbol{\Upsilon}}$ is badly biased for $\boldsymbol{\Upsilon}$, or when $\mathsf{E}[\mathbf{x}_{T+1} \mid \mathbf{x}_T] = \boldsymbol{\Psi} \boldsymbol{x}_T$ but $\mathsf{E}[\mathbf{x}_{T+h} \mid \mathbf{x}_T] \neq \boldsymbol{\Psi}^h \boldsymbol{x}_T$. However, they remarked also that, in stationary processes, misspecification of the DGP is not sufficient to advocate the use of DMS, since $\widehat{\boldsymbol{\Upsilon}}$ is the OLS and $\widetilde{\boldsymbol{\Upsilon}}_h$ converges towards the unconditional expectation with $\boldsymbol{\Upsilon}^h$ tending to zero as h increases. Hence, increasing divergence between $\widehat{\boldsymbol{\Upsilon}}$ and $\widetilde{\boldsymbol{\Upsilon}}_h$ is unlikely. Moreover, DMS is inefficient in small samples, so that if $\boldsymbol{\epsilon}_t \sim \mathsf{IN}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$, biases are unlikely to be enough for a gain to appear. Thus, Clements and Hendry noted that if $\boldsymbol{\epsilon}_t$ follows a negative moving average, there may be some potential for DMS. They derived a taxonomy of forecast errors and showed that the only terms in common for both methods were those of error accumulation, namely $\sum_{i=0}^{h-1} \boldsymbol{\Upsilon}^i \boldsymbol{\epsilon}_{T+h-i}$ in (21). Simulating small sample estimation biases, they showed, for several stationary values – 0, 0.4 and 0.8 – of the autoregressive coefficient in a univariate AR(1) process without intercept, that for sample sizes ranging from 10 to 100 the two-step-ahead DMS does not yield better estimates of the powered coefficient than the squared IMS.

This result is specific to finite samples as Chevillon and Hendry (2005) showed. They estimated (20), with an additional drift, by OLS and (21), with a drift also, by the generalized method of moments with a heteroscedasticity and autocorrelation consistent covariance matrix estimator. DMS was asymptotically more efficient than IMS in the case of stationary processes with a positive slope. Indeed, in the univariate case, denoting by $\widehat{e}_h$ and $\widetilde{e}_h$ the IMS and DMS forecast errors at horizon h using generalized method of moments estimation, and ρ the slope coefficient, these authors showed that

$$h \left(\mathsf{E}\left[\widehat{e}_h^2\right] / \mathsf{E}\left[\widetilde{e}_h^2\right] - 1 \right) \underset{h\to\infty}{\to} \frac{2\rho}{(1-\rho^2)(2-\rho)}$$

the latter being of the same sign as ρ, as long as $|\rho| < 1$. In the case of integrated processes, this result collapses and IMS always dominates DMS.

A Monte Carlo analysis by Clements and Hendry (1996) of the forecasts from the 'non-seasonal Holt–Winters model' illustrated the relative behaviours of the IMS and DMS techniques in their framework. The data were generated by the sum of unobserved components for the trend, intercept and irregular elements:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_t &= \mu_{t-1} + \beta_t + \delta_{1t} \\ \beta_t &= \beta_{t-1} + \delta_{2t} \end{aligned}$$

The disturbances ε_t, δ_{1t} and δ_{2t} are assumed to be normally distributed and independent through time and from one another (at all lags and leads), with zero means and variances, respectively, σ_ε^2, $\sigma_{\delta_1}^2$ and $\sigma_{\delta_2}^2$. This model can be reduced to an ARIMA(0, 2, 2) – or restrictions thereof – with or without a drift, or a deterministic trend. It is also possible to allow for stationary autoregressive roots,

$$(1 - \tau_1 L)(1 - \tau_2 L) y_t = \delta_{2t} + (1 - L)\delta_{1t}$$

The authors used six AR forecasting models: AR(2) models in levels or differences, with or without an imposed unit root, with or without an intercept. The main results are that DMS and IMS are somewhat equivalent when the model either estimates the unit root or neglects moving average (MA) components. However, when these two effects are present, there is a gain for DMS (seemingly increasing with the horizon) unless the MA term is effectively cancelled by an AR root, or when the model is underparameterized for the DGP. The forecasts, using the pseudo-true values of the parameters for the DGP considered, allowed the separation of the effects of model misspecification and estimation uncertainty. In general, the misspecification effects are constant or even decrease with the horizon, and multi-step forecasts can be more accurate in the very near future. In terms of estimation, DMS is more accurate when one (or two) unit root is present in the DGP but not imposed in the model, and in the presence of omitted MA errors (the conjunction of both seems necessary, as opposed to either alone). A significant gain is present also when an intercept is estimated in conjunction with the other two effects, especially when the FGP is an AR(1), for which IMS fares badly. These results help explain why Stoica and Nehorai (1989) found that a model close to an ARIMA(0, 1, 2) approximated by an AR(1) led to improved forecasting performance when using DMS but not when the FGP is an AR(6).

Clements and Hendry then focused on the driftless ARIMA(0, 1, 1) DGP, where the MA component is omitted in the forecasting models and the unit root is estimated. They used four estimators for the hth power of the slope in $y_t = \rho y_{t-1} + \epsilon_t$ and $\epsilon_t = \zeta_t + \theta\zeta_{t-1}$, where $\zeta_t \sim \mathsf{IN}(0, \sigma_\zeta^2)$:

$$(\widehat{\rho}_{1\mathrm{S}})^h = \left(\frac{\sum y_t y_{t-1}}{\sum_t y_{t-1}^2}\right)^h \quad \text{and} \quad \widetilde{\rho}_{\mathrm{DMS}_h} = \frac{\sum y_t y_{t-h}}{\sum_t y_{t-h}^2}$$

$$(\widehat{\rho}_{\mathrm{IV}})^h = \left(\frac{\sum y_t y_{t-2}}{\sum_t y_{t-1} y_{t-2}}\right)^h \quad \text{and} \quad \widetilde{\rho}_{\mathrm{IVDMS}_h} = \frac{\sum y_t y_{t-h-1}}{\sum_t y_{t-h} y_{t-h-1}}$$

They showed that

$$\begin{aligned}
T\left((\widehat{\rho}_{1\mathrm{S}})^h - 1\right) &\Rightarrow \left(\int_0^1 W(r)^2 \mathrm{d}r\right)^{-1} h\left[\frac{1}{2}(W(1)^2-1) + \frac{\theta}{(1+\theta)^2}\right]\\
T\left((\widehat{\rho}_{\mathrm{IV}})^h - 1\right) &\Rightarrow \left(\int_0^1 W(r)^2 \mathrm{d}r\right)^{-1} \frac{h}{2}(W(1)^2-1)\\
T(\widetilde{\rho}_{\mathrm{DMS}h} - 1) &\Rightarrow \left(\int_0^1 W(r)^2 \mathrm{d}r\right)^{-1} h\left[\frac{1}{2}(W(1)^2-1) + \frac{\theta}{h\,(1+\theta)^2}\right]\\
T(\widetilde{\rho}_{\mathrm{IVDMS}h} - 1) &\Rightarrow \left(\int_0^1 W(r)^2 \mathrm{d}r\right)^{-1} \frac{h}{2}(W(1)^2-1)
\end{aligned}$$

and provided small sample approximations of the distributions. The leftward non-centrality of IMS therefore increases with h, whereas that of DMS does not. The instrumental estimators seem better. Simulations illustrate these results.

This framework was also analysed in Chevillon and Hendry (2005) who now allowed for a drift in the random walk. This induces the presence of a deterministic trend which asymptotically dominates estimation, yielding the same asymptotic accuracy for both the methods. In a finite sample, though, disparities appear: DMS is more accurate when the drift is 'small' compared to the variance of the disturbances and when the latter exhibit negative serial correlation. Introducing the concept of a 'weak' trend whereby the drift coefficient vanishes to zero asymptotically at the rate of $\mathrm{O}(T^{-1/2})$, Chevillon (2005) derived asymptotic distributions where he allowed for both the stochastic and deterministic trends to have an impact on estimation. The local asymptotic model he used is

$$y_t = \left(\sum_{i=0}^{h-1} \rho^i\right)\tau_T + \rho^h y_{t-h} + \varepsilon_t \quad \text{for } h \geq 1$$

where $\tau_T = \psi/\sqrt{T}$, $\mathrm{Var}[\varepsilon_t] = \sigma_\varepsilon$, $\sigma^2 = \lim_{T\to\infty} T^{-1}\mathsf{E}[\sum_{t=1}^T \varepsilon_t]$ and the autocovariance function of $\{\varepsilon_t\}$ is given by $\gamma_i^{(\varepsilon)}$. The resulting IMS, $(\widehat{\tau}_T^{\{h\}}, \widehat{\rho}_T^h)$, and DMS, $(\widetilde{\tau}_{h,T}, \widetilde{\rho}_{h,T})$, estimators are such that

$$\begin{aligned}
&\begin{bmatrix}\sqrt{T}(\widetilde{\tau}_{h,T} - \tau_{h,T})\\ T(\widetilde{\rho}_{h,T} - 1)\end{bmatrix} - \begin{bmatrix}\sqrt{T}(\widehat{\tau}_T^{\{h\}} - \tau_{h,T})\\ T\left(\widehat{\rho}_T^h - 1\right)\end{bmatrix}\\
&\quad\Rightarrow \frac{\sum_{i=1}^{h-1}(h-i)\,\gamma_i^{(\varepsilon)}}{\int_0^1 [K_{\psi,\phi}(r)]^2 \mathrm{d}r - \left(\int_0^1 K_{\psi,\phi}(r)\mathrm{d}r\right)^2}\left[\int_0^1 K_{\psi,\phi}(r)\mathrm{d}r - 1\right] \qquad (22)
\end{aligned}$$

where $K_{\psi,\phi}$ is a drifting Ornstein–Uhlenbeck process defined as

$$K_{\psi,\phi}(r) = \psi f_\phi(r) + \sigma\int_0^r \mathrm{e}^{\phi(r-s)}\mathrm{d}W(s)$$

with $W(r)$ a Wiener process on [0, 1] and

$$f_\phi(\cdot)\colon r \to \frac{\mathrm{e}^{\phi r}-1}{\phi} \quad \text{if } \phi \neq 0 \quad \text{and } f_0(r) = r \tag{23}$$

The difference between the two types of estimators is a function of $\|\psi, \sigma\|$. In turn, this translates in the forecast errors which the author showed to be complex functions of the forecast horizon and parameters. Analysis of the distributions and Monte Carlo simulation prove that the weak trend framework accurately represents the finite sample behaviours and that it is the ratio ψ/σ that matters in this context. Yet, equation (22) shows that negative autocorrelation of the innovations will improve the estimation of the (underestimated) direct multi-step slope coefficient compared to iterating. As the covariance of the intercept and slope biases has a sign opposite that of τ, an improvement in slope estimation tends to lead to a similar result for the drift. Furthermore, if $\gamma_i^{(\varepsilon)}$ is zero for $i \geq p$, then whichever method 'wins' at horizon p does so increasingly for higher horizons. In the same paper, the author obtained the actual distributions of the forecast error as expansions in the order of the ratio h/T. He showed that estimation of the multi-step parameters in terms of bias and variance is key at small horizons. The complexity of interactions prevented him from conclusions at larger horizons.

Deterministic misspecification has also been shown to benefit direct multi-step estimation. As mentioned in Chevillon and Hendry (2005), occasional breaks in the level of a trending process can generate serial correlation of the residuals from a constant parameter model and lead to the cases studied by these authors. Chevillon (2006) also analysed the influence of recent unnoticed breaks. He showed that DMS is more efficient at estimating the dynamic properties relevant for forecasting and that the potential occurrence of deterministic shocks hence advocates using direct methods. This aspect is confirmed in an empirical forecasting exercise for the South African GDP over 1973–2000 where a multi-step method designed by Aron and Muellbauer (2002) and variants thereof beat all of 30 rival techniques.

In extensive empirical studies, Kang (2003) and Marcellino *et al.* (2006) found little improvement using DMS, but they carried out estimation on stationary time series (e.g. first differences) and evaluated forecasts on the levels. Non-stationarity therefore appears key.

8.1 *Discussion*

These authors confirmed, with their Monte Carlo, and proved analytically what some previous authors had found in specific cases, namely that estimated unit roots, structural breaks and omitted negative residual autocorrelation are key to the success of DMS forecasting. As opposed to some other authors, they used as a DMS model the projection of the variable onto the space spanned by its lags at and beyond h: it is the same autoregressive dynamics which is estimated. Their simulations also shed light on earlier results. The influence of small drifts is shown and it is seen that in general DMS is to be preferred when the data are – stochastically or deterministically – non-stationary or when the available sample is too small for reliable inference.

9. Testing for Multi-step Forecast Accuracy

Two very recent papers focus on testing multi-step forecast accuracy. This is a different issue from those of the design of FGPs and evaluation of the factors that influence the accuracy but is still mostly unresolved. What matters in this context is whether one model can *ex ante* or *ex post* be tested to yield superior forecasts. The criterion with the most widespread use for evaluating forecast accuracy is the minimum mean square forecast error (MMSFE) conditional on the available observation. This is defined in the case of multi-step forecasts as

$$\text{MSFE}_h \equiv \mathsf{E}[\mathbf{e}_{T+h}\mathbf{e}'_{T+h}|\mathbf{X}_T] = \mathsf{V}[\mathbf{e}_{T+h}|\mathbf{X}_T] + \mathsf{E}[\mathbf{e}_{T+h}|\mathbf{X}_T]\mathsf{E}[\mathbf{e}'_{T+h}|\mathbf{X}_T]$$

where $\mathbf{e}_{T+h}$ is the forecast error. Clements and Hendry (1993) showed analytically that this criterion is not invariant to 'non-singular, scale-preserving, linear transformations for which the associated model class is invariant'. It may happen that bijective mappings of the variables and models (e.g. whether levels or difference are considered) yield conflicting rankings of the forecasting techniques. Alternative criteria overcome this weakness. For instance, the general forecast error second moment (GFESM) (see Clements and Hendry, 1993) is related to the concept of predictive likelihood presented in Bjønstad (1990), and defined as

$$\text{GFESM}_h = |\mathbf{G}_h|$$

where $\mathbf{G}_h = \mathsf{E}[\mathbf{E}\mathbf{E}'|\mathbf{X}_T]$ and $\mathbf{E} = [\mathbf{e}'_{T+1}, \ldots, \mathbf{e}'_{T+h}]'$. GFESM$_h$ is invariant to linear transforms. This is close to the estimation criterion in (8), where several forecasting horizons are considered together (also in Weiss, 1991; Liu, 1996). Since the idea underlying multi-step estimation lies in the re-estimation of the parameters at each horizon, using a deliberately non-congruent model, it seems unavoidable that forecast accuracy should rely strongly on the dynamic properties of the variables to be forecast. Hence, the flaws exhibited by the MSFE do not matter in direct multi-step forecasting and its computational ease makes such a criterion all the more attractive.

An additional difficulty associated with the MSFE is simply that it may not exist. This was shown in an AR(1) context by Hoque *et al.* (1988) when the forecast horizon is too large relative to the size of the available sample. However, as pointed out by Ericsson and Marquez (1998), even if the variance of the forecast errors is infinite in small samples, sensible results can be obtained in practice and they, effectively, can be seen as Nagar expansions (see Sargan, 1982). Indeed, the observed – if at all – low frequency of large outliers provides a truncated approximation to the tails of the distribution.

The *ex ante* MSFE is a statistic which, when derived analytically or by simulation, reflects some known properties of the forecast error distribution. In theoretical studies, many authors assume that the processes used for estimation and forecasting are independent. This was fervently criticized by Ing (2004) who showed that when taking estimation uncertainty in the design of the forecasting model, and hence evaluating accuracy using the *unconditional* MSFE, ranking reversals can occur. This is why this author uses a related criterion – accumulated multi-step prediction error – in the design of his models (this constitutes an empirical mean of the squares of recursive prediction errors).

Unfortunately, for *ex ante* MSFE dominance to constitute a valid reference for actual superior forecast accuracy, the DGP need not undergo structural breaks and some of its properties must be known. With these assumptions, though, some inference is possible: in a recent paper, Haywood and Tunnicliffe-Wilson (2004) developed a score test for multi-step forecasting accuracy. This allowed comparison of an IMS model estimated by minimizing the squares of in-sample one-step residuals with the same functional form but with parameters estimated by the in-sample multi-step residuals (parametric DMS as in Section 4). They suggested using as a statistic the derivative of the h-step IMS MSFE with respect to the model parameters (estimated in the frequency domain). This presents the advantage of allowing for some forms of non-stationarity. Failure to reject the test implies that parametric DMS yields no improved forecast accuracy upon IMS.

In practice, when the MSFE is computed *ex post* from empirical data, it needs to be complemented with assumptions about the actual distributions of the forecast errors so that comparisons across models can be made. One of the most common tests for equality of MSFEs was derived by Diebold and Mariano (1995) and Harvey *et al.* (1997). This is a version of a Hausman test (see Hausman, 1978) where, letting two different procedures produce a pair of h-step-ahead forecast errors $\widehat{\mathbf{e}}_{t+h|t}$ and $\widetilde{\mathbf{e}}_{t+h|t}$, for $t = 1, \ldots, T$, and if the forecast is to be judged on some criterion $\mathcal{C}_F(\mathbf{e})$, then the null hypothesis of equal expected forecast accuracy is

$$\mathsf{H}_0\colon \mathsf{E}\left[\mathcal{C}_F\left(\widehat{\mathbf{e}}_{t+h|t}\right) - \mathcal{C}_F\left(\widetilde{\mathbf{e}}_{t+h|t}\right)\right] = \mathbf{0}$$

The sample mean of $\mathbf{c}_t = \mathcal{C}_F\left(\widehat{\mathbf{e}}_{t+h|t}\right) - \mathcal{C}_F\left(\widetilde{\mathbf{e}}_{t+h|t}\right)$, excluding an initializing subsample of T_1 observations, is denoted by $\overline{\mathbf{c}}$ so that for a well-specified model the series of forecast errors should follow a vector moving average (VMA($h-1$)) and $\overline{\mathbf{c}}$ should exhibit autocorrelation up to order $h-1$ (this assumption can be readily extended to allow for a VMA(q), with $q \geq h-1$). Defining γ_k the sequence of autocovariance matrices of $\mathbf{c}_t$, estimated by

$$\widehat{\gamma}_k = \frac{1}{T - T_1 + 1} \sum_{t=T_1+k}^{T} (\mathbf{c}_t - \overline{\mathbf{c}})(\mathbf{c}_{t-k} - \overline{\mathbf{c}})'$$

then asymptotically

$$\widehat{\mathsf{V}}[\overline{\mathbf{c}}] = \frac{1}{T - T_1 + 1} \left(\widehat{\gamma}_0 + \sum_{k=1}^{h-1} \left[\widehat{\gamma}_k + \widehat{\gamma}_k' \right] \right) \underset{T\to\infty}{\to} \mathsf{V}[\overline{\mathbf{c}}]$$

The Diebold–Mariano test statistic is defined as $\mathrm{DM}_T = [\widehat{\mathsf{V}}[\overline{\mathbf{c}}]]^{-1/2}\,\overline{\mathbf{c}}$, and, under the null hypothesis,

$$\mathrm{DM}_T \underset{T\to\infty}{\overset{L}{\to}} \mathsf{N}\left(\mathbf{0}_{n\times 1}, \mathbf{I}_n\right)$$

$\widehat{\mathsf{V}}[\overline{\mathbf{c}}]$ is a heteroscedasticity and autocorrelation consistent estimator of the variance–covariance matrix of $\overline{\mathbf{c}}$: it is hence subject to the associated difficulties. It so happens that, in finite samples, $\widehat{\mathsf{V}}[\overline{\mathbf{c}}]$ may not be positive semi-definite and be a biased estimator of $\mathsf{V}[\overline{\mathbf{c}}]$; this led Harvey *et al.* (1997) to suggest a *modified* DM test where the estimator of $\mathsf{V}[\overline{\mathbf{c}}]$ was corrected, and in Harvey *et al.* (1998) to provide a test for multi-step forecast encompassing where $\widehat{\mathbf{e}}_{t+h|t}$ is then compared to $(\widehat{\mathbf{e}}_{t+h|t} - \widetilde{\mathbf{e}}_{t+h|t})$.

Clark and McCracken (2005) considered the asymptotic and finite sample properties of the previous tests of equal multi-step forecast accuracy and encompassing applied to nested models as the distribution of the DM statistics differed in this setting. In their framework the two competing direct multi-step forecasts are

$$\begin{aligned} \mathrm{M}_1\colon\ y_{t+h} &= \boldsymbol{\beta}_1' \mathbf{x}_{1,t} + \varepsilon_{1,t+h} \\ \mathrm{M}_2\colon\ y_{t+h} &= \boldsymbol{\beta}_1' \mathbf{x}_{1,t} + \boldsymbol{\beta}_2' \mathbf{x}_{2,t} + \varepsilon_{2,t+h} \end{aligned}$$

Recursive estimation is carried out on a sample of R observations. This leads to $P - h + 1$ forecasts where $R + P = T$. These authors also analysed the F-test equivalents to the t above for stationary processes. The multi-step tests are unfortunately non-similar as their distributions depend on nuisance parameters. Yet, their estimation framework which closely relates to that of Ing's APE (see Section 6), provided the distribution of recursive *ex ante* forecast accuracy tests which, when complemented with some stability (no posterior breaks...) and specification assumptions, allowed the modeller to carry out *ex post* testing strategies for multi-step forecast accuracy.

9.1 *Discussion*

Testing for improved multi-step forecasting is an important issue which would allow practitioners to gauge *ex ante* whether IMS or DMS ought to be used in a specific case. Such tests, when available, will only be conditional on some stability in the DGP so that in-sample testing provides a reliable guide to out-of-sample forecasting. No general results are yet available but the two papers by Clark and McCracken (2005) and Haywood and Tunnicliffe-Wilson (2004) led paths in two different but essential directions. The first authors were interested in obtaining a reliable statistic for the design of the DMS model; the second in testing from the model parameters whether IMS will beat DMS.

10. Direct Multi-step Estimation and Forecasting

10.1 *Design of Forecast Estimators*

In this section, we provide a general definition for the two types of forecasts which we have studied so far, namely the iterated one-step ahead (IMS$_h$) and direct h-step (DMS$_h$). We borrow a framework for the design of forecast estimators from Ericsson and Marquez (1998) and extend it to allow for multi-step estimation. Here, the modeller is interested in n endogenous variables, $\mathbf{x}$, and assumes that they depend on their lagged values, up to some $p \geq 0$, on some weakly exogenous – which they actually may or may not be – variables $\mathbf{z}$ and on some vector of c parameters $\boldsymbol{\varphi}$. The model specifies some error process $\{\boldsymbol{\epsilon}_t\}$ – the distribution thereof may depend on $\boldsymbol{\varphi}$ and exhibit any form of autocorrelation, heteroscedasticity or non-stationarity –

and is assumed to be valid over a sample of size $T+H$, so that there exists an n-vector function $\mathbf{f}(\cdot)$, such that

$$\mathbf{f}(\mathbf{x}_t, \ldots, \mathbf{x}_{t-p}, \mathbf{z}_t, \boldsymbol{\varphi}, \boldsymbol{\epsilon}_t) = \mathbf{0} \quad \text{for } t = p, \ldots, T, \ldots, T + H \tag{24}$$

The sample is split into two: estimation is conducted over the first T observations and this is used to forecast the remaining H. Equation (24) describes an open model and it is convenient to transform it in a reduced closed form, solving it for $\mathbf{x}_t$. We change the time subscript t to $T+i$, and assume – under mild conditions, among which linearity of $\mathbf{f}(\cdot)$ is most common – that there exists a suitable transform of $\mathbf{f}(\cdot)$, denoted by $\mathbf{g}(\cdot)$, such that positive values of i represent the dates for which we wish to obtain forecasts in

$$\mathbf{x}_{T+i} = \mathbf{g}(\mathbf{x}_{T+i-1}, \ldots, \mathbf{x}_{T+i-p}, \mathbf{z}_{T+i}, \boldsymbol{\varphi}, \boldsymbol{\epsilon}_{T+i}) \quad \text{for } i = p - T, \ldots, -1, 0, 1, \ldots, H \tag{25}$$

This framework is quite general, and it may be the case that specific models should be restrictions thereof. For instance, if $n = 1$, $\mathbf{g}(\cdot)$ reduces to a single equation; it may also be non-linear and the model could be static – if $p = 0$ – or exclude exogenous variables.

For forecasting at horizons $i > 1$, there is a need for assumptions about the vector $\mathbf{z}_t$: either it is assumed to be strongly exogenous and it is possible to obtain conditional forecasts (see Engle *et al.*, 1983), or a model for its behaviour is used, and in fact $\mathbf{z}$ is incorporated in $\mathbf{x}$. The forecasts are defined by their horizon, i, the actual variable of interest – which can be a transform of $\mathbf{x}_{T+i}$ – and the specification of its distribution, as given here by (25).[2] What values of $\mathbf{x}_{T+i-1}, \ldots, \mathbf{x}_{T+i-p}$, $\mathbf{z}_{T+i}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\epsilon}_{T+i}$ are used in forecasting affects the outcome. For instance, the one-step-ahead forecast, from an end-of-sample forecast origin at T, is obtained when the actual values of $\mathbf{x}_T, \ldots, \mathbf{x}_{T-p}$ are used in $\mathbf{g}(\cdot)$. And then, for $i > 1$, replacing $\mathbf{x}_{T+i-1}$ in the equation with its corresponding forecast leads to iterated one-step-ahead forecasts (IMS) and (25), specifying the parameters $\boldsymbol{\varphi}$ and the distributions of the disturbances, is thus the corresponding FGP.

Alternatively, it is possible to directly estimate the DGP h steps ahead, for a fixed $h > 1$, using a transformed representation of (24). We let $\mathbf{k}_h(\cdot)$ denote a suitable transform of $\mathbf{f}(\cdot)$ – possibly including some composition – such that

$$\begin{aligned} \mathbf{x}_{T+i} &= \mathbf{k}_h(\mathbf{x}_{T+i-h}, \ldots, \mathbf{x}_{T+i-h-p+1}, \mathbf{w}_{T+i}, \boldsymbol{\phi}_h, \boldsymbol{\nu}_{h,T+i}) \\ &\text{for } i = p - 1 + h - T, \ldots, -1, 0, 1, \ldots, H \end{aligned} \tag{26}$$

where $\boldsymbol{\phi}_h$, a c-vector of parameters, and $\boldsymbol{\nu}_{h,t}$, an n-vector of disturbances, are re-parameterizations of $\boldsymbol{\varphi}$ and $\boldsymbol{\epsilon}_t$. The r-vector $\mathbf{w}_{t+i}$ is assumed to be a transform of $\{\mathbf{z}_t\}$ which achieves a property of strong exogeneity for the parameters of (26), namely $\boldsymbol{\phi}_h$. The forecasts $\{\widetilde{\mathbf{x}}_{T+i,h}; i = 1, \ldots, H\}$ obtained using $\mathbf{k}_h(\cdot)$ are the multi-step forecasts of $\{\mathbf{x}_{T+i}; i = 1, \ldots, H\}$, using dynamic – or direct – estimation, the h-step DMS forecasts, generated by the DMS FGP (26). The exogeneity status of $\mathbf{z}_{T+i}$ and $\mathbf{w}_{t+i}$ may be misspecified in practice; additional uncertainty is generated when forecasts are used instead of the true realized values, especially given that their FGPs may not coincide with their DGPs.

If the modeller knew with certainty the DGP and it coincided with her model (24), then both IMS and DMS FGPs would provide the same forecasts. In practice, unfortunately, (24), (25) and (26) would have to be estimated and depending on which methods are used for this purpose the estimated parameters[3] $\widehat{\varphi}$ and $\widetilde{\phi}_h$ will lead to different forecasts. The inter-dependence between *estimation* and *forecasting* is therefore intrinsic to the concept of multi-step forecasting. Before presenting a forecast error taxonomy which will help us in assessing the factors beneficial to DMS, we first focus on the issue of selecting the regressors.

10.2 *Model Choice*

As pointed out by a referee, most of the existing results concern or focus mostly on the univariate context, and additional possibilities exist in a multivariate framework that directly affect the relative choice between IMS and DMS. In particular, a difficulty associated with model selection which is of importance for multi-step forecasting concerns the choice of regressors, $\mathbf{z}_t$ in (24). In most models, the set of variables $\mathbf{z}_t$ would be weakly exogenous for φ, thus enabling contemporaneous estimation and forecasting of $\mathbf{x}_t$ as in (25) which, for simplicity, we assume linear and without lags of the endogenous variables:

$$\mathbf{x}_t = \varphi' \mathbf{z}_t + \epsilon_t \tag{27}$$

A forecast of $\mathbf{x}_{T+h}$ from the end of the sample T is therefore obtained using

$$\mathsf{E}\left[\mathbf{x}_{T+h}\right] = \mathsf{E}[\varphi' \mathbf{z}_{T+h}]$$

and necessitates that the modeller produces a forecast for $\mathbf{z}_{T+h}$. If the regressors are stochastic, $\mathbf{z}_{T+h}$ needs to be approximated by its forecast $\widehat{\mathbf{z}}_{T+h}$ which itself can only be generated when the regressors are strongly exogenous for φ so that $\mathbf{z}_{T+h}$ does not depend on $\{\mathbf{x}_t\}$ for $t > 1$. If strong exogeneity is not a reliable hypothesis, then in-sample congruent modelling such as (27) does not preclude multi-step forecast failure, whether or not lags of $\mathbf{x}_t$ matter. Another, less stringent property may be of use in this context: *multi-step weak exogeneity* of the regressor ($\mathbf{z}_t$) for the *multi-step* parameter of interest (χ_h), for a given h. This can be defined, in terms of the variables $\mathbf{y}_t = (\mathbf{x}_t, \mathbf{z}_t)$, as

$$\text{(1):}\quad \mathbf{f}_{h,\mathbf{y}}\left(\mathbf{x}_t, \mathbf{z}_t \,|\, \{\mathbf{y}_i\}_{t-h}^{1}, \psi_h\right) = \\ \mathbf{f}_{h,\mathbf{y}|\mathbf{z}}\left(\mathbf{x}_t, \mathbf{z}_{t-h} \,|\, \{\mathbf{y}_i\}_{t-h}^{1}, \varphi_h\right) \times \mathbf{f}_{h,\mathbf{z},h}\left(\{\mathbf{z}_i\}_{t}^{t-h} \middle| \{\mathbf{z}_i\}_{t-h}^{1}, \phi_h\right)$$

(2a): χ_h is a subset of ψ_h and a function of φ_h alone and

(2b): φ_h and ϕ_h are variation free

where $\mathbf{f}_{(\cdot,\cdot)}$ represents the density function and $t = 1, \ldots, T + h$. When (1), (2a) and (2b) are satisfied, then direct multi-step estimation is possible and a valid method to obtain forecasts of $\mathbf{x}_{T+h}$ conditional on $\{\mathbf{x}_t, \mathbf{z}_t\}_T^1$ such as in

$$\mathbf{x}_{t+h} = \varphi_h' \mathbf{z}_t + \mathbf{w}_{t+h}$$

In this context DMS is the only valid procedure, or endogenously modelling $\mathbf{z}_t$ is necessary. DMS thus constitutes a way to alleviate some of the potential risks associated with forecasting $\mathbf{z}_t$. Yet, when the DGP is subject to breaks no method can guarantee an accurate forecast. Clements and Hendry (2004) showed that extraneous shocks affecting the weakly exogenous regressors constitute a major source of forecast failure. One option is to leave some regressors out altogether so that IMS is feasible. A well-known result of omitted regressors is that provided these are stationary or that their out-of-sample expectations and variances match their in-sample values, they do not lead to systematic forecast biases and the MSFE is close to that anticipated by the in-sample estimates (see Clements and Hendry, 1999, pp. 99–102). In a macroeconomic model following an equilibrium correction mechanism the regressors are stationary and hence their omission should not generate systematic forecast failure, unless the cointegrating vectors are misspecified so that a stochastic trend is induced.

In practice, multi-step weak exogeneity may prove difficult to establish and the choice remains as to which regressors deliberately not to model. Clements and Hendry (2005) presented a criterion when a change of collinearity among the $\mathbf{z}_t$ variables can potentially happen. They showed that a variable should be included in a forecasting model when the square of the t-test statistic for its parameter (under the null of insignificance) has a non-centrality $\tau^2 > 1$. This criterion also matters for DMS when there are more potential regressors than observations so that factor analysis is carried out. Iterated multi-step is not possible in this case. Hendry and Hubrich (2005) showed that when multi-step forecasting is desired, the criterion for including a dynamic regressor (for instance, the first lag of the endogenous variable) depends on the forecast horizon ($\tau^2_{\rho=0} > h^2$ in their example where ρ is the slope in a first-order autoregression).

The issue of model choice for multi-step forecasting and of the practical benefits from direct methods is still unresolved. Yet, there exist many cases when this technique proves the only option or when model misspecification cannot easily be corrected to provide iterated forecasts. We present below a forecast error taxonomy which aims at pointing out the context most beneficial to DMS.

10.3 *A General Forecast Error Taxonomy*

We now borrow from Clements and Hendry who suggested in Clements and Hendry (1998b) and Hendry (2000) a general forecast error taxonomy which helps in assessing the advantages of multi-step estimation. We use the framework presented above but for ease of exposition modify it slightly. Notice that in (24) $\mathbf{f}_{\mathbf{x}}(\cdot)$, if it represents the true DGP, provides the – potentially time-dependent – joint density of $\mathbf{x}_t$ at time t, conditional on $\mathbf{X}_{t-1}^{t-p} = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-p})$, and $\mathbf{z}_t$. Assume, without loss of generality, that $\{\mathbf{z}_t\}$ contains only deterministic factors – such as intercepts, trends and indicators – and that all stochastic variables are included in $\{\mathbf{x}_t\}$. As previously, it is desired to forecast $\mathbf{x}_{T+h}$, or perhaps a function thereof (e.g. if $\mathbf{z}_t$ originally contained stochastic variables), over horizons $h = 1, \ldots, H$, from a forecast origin at T. Now, the dynamic model does not coincide with the DGP and it specifies the distribution of $\mathbf{x}_t$ conditional on $\mathbf{X}_{t-1}^{t-r}$, with lag length r, deterministic components $\mathbf{d}_t$ and implicit stochastic specification defined by its parameters ψ_t. This model is fitted over the sample $t = 0, \ldots, T$, so that parameter estimates are a function of the observations, represented by

$$\widehat{\psi}_T = \boldsymbol{\Psi}_T\left(\widetilde{\mathbf{X}}_T^0, \mathbf{D}_T^0\right) \tag{28}$$

where $\widetilde{\mathbf{X}}$ denotes the measured data and, as before, $\mathbf{D}_t^0 = (\mathbf{d}_t, \ldots, \mathbf{d}_0)$. A sequence of forecasts $\{\widehat{\mathbf{x}}_{T+h|T}\}$ is produced as a result. The subscript on $\widehat{\psi}$ in (28) denotes the influence of the sample size. Let $\psi_T^e = \mathsf{E}_T[\widehat{\psi}_T]$, where it exists. Because the underlying densities may be changing over time, all expectation operators must be time dated. Future values of the stochastic variables are unknown, but those of deterministic variables are known; there therefore exists a function $\mathbf{g}_h(\cdot)$ such that

$$\widehat{\mathbf{x}}_{T+h|T} = \mathbf{g}_h\left(\widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \widehat{\psi}_T\right) \tag{29}$$

The corresponding h-step-ahead expected forecast error is thus the expected value of $\mathbf{e}_{T+h|T} = \mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T}$, and is given by

$$\mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right]$$

where the actual values of the deterministic factors over the forecast period (including any deterministic shifts) are denoted by $\{\mathbf{Z}^*\}_{T+h}^{T+1}$ and $\{\mathbf{Z}^*\}_{T+h}^0 = [\{\mathbf{Z}^*\}_{T+h}^{T+1}, \mathbf{Z}_T^0]$; and the expectation operator is dated $T + h$ to take account of the model specification of the deterministic components between $T + 1$ and $T + h$. The expectation of $\widehat{\mathbf{x}}_{T+h|T}$, conditional on the information available at T and on the assumptions made about the interval $T + 1, \ldots, T + h$, is the *model-induced* conditional expectation. Define, from an origin T, the h-step disturbance

$$\varepsilon_{T+h|T} = \mathbf{x}_{T+h} - \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right] \tag{30}$$

By construction, $\mathsf{E}_{T+h}[\varepsilon_{T+h|T} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0] = 0$ and $\varepsilon_{T+h|T}$ is therefore an innovation against all available information. However, even for correctly observed sample data, it is not, in general, the case that

$$\mathsf{E}_{T+h}\left[\mathbf{e}_{T+h|T} \mid \mathbf{X}_T^0, \left\{\mathbf{Z}^*\right\}_{T+h}^0\right] = 0$$

as we now show.

Using (30), the forecast error $\mathbf{e}_{T+h|T} = \mathbf{x}_{T+h} - \widehat{\mathbf{x}}_{T+h|T}$ from the model based on (29) can be decomposed as

$$\mathbf{e}_{T+h|T} = \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \{\mathbf{Z}^*\}_{T+h}^0\right] - \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] \tag{ia}$$

$$+ \mathsf{E}_{T+h}\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] - \mathsf{E}_T\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] \tag{ib}$$

$$+\,\mathsf{E}_T\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{Z}_{T+h}^0\right] - \mathsf{E}_T\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^0, \mathbf{D}_{T+h}^0\right] \tag{iia}$$

$$+\,\mathsf{E}_T\left[\mathbf{x}_{T+h} \mid \mathbf{X}_T^1, \mathbf{D}_{T+h}^1\right] - \mathsf{E}_T\left[\widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \psi_T^e\right] \tag{iib}$$

$$+\,\mathsf{E}_T\left[\widehat{\mathbf{x}}_{T+h|T} \mid \mathbf{X}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \psi_T^e\right] - \mathsf{E}_T\left[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \psi_T^e\right] \tag{iii}$$

$$+\,\mathsf{E}_T\left[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \psi_T^e\right] - \widehat{\mathbf{x}}_{T+h|T} \tag{iv}$$

$$+\,\varepsilon_{T+h|T} \tag{v}$$

The first two rows arise from structural change affecting deterministic (ia) and stochastic (ib) components respectively; the third and fourth, (iia) and (iib), from model misspecification decomposed by deterministic and stochastic elements; the fifth (iii) from forecast origin inaccuracy; (iv) represents estimation uncertainty; and the last row, (v), is the unpredictable stochastic component.

When $\{\mathbf{Z}^*\}_{T+h}^0 = \mathbf{Z}_{T+h}^0$ (i.e. in the absence of deterministic shifts), then (ia) is zero; and, in general, the converse holds, that (ia) being zero entails no deterministic shifts. When $\mathsf{E}_{T+h}[\cdot] = \mathsf{E}_T[\cdot]$ (so that there are no stochastic breaks), (ib) is zero; but (ib) can be zero despite stochastic breaks, provided that these do not indirectly alter deterministic terms. When the deterministic terms in the model are correctly specified, so that $\mathbf{Z}_{T+h}^0 = \mathbf{D}_{T+h}^0$, then (iia) is zero, and again the converse seems to hold. In the case of correct stochastic specification, so that ψ_T^e summarizes the effects of $\mathbf{X}_T^1$, then (iib) is zero; but now the converse is not true: (iib) can be zero in seriously misspecified models. Next, when the data are accurate (especially at the forecast origin), so that $\mathbf{X} = \widetilde{\mathbf{X}}$, (iii) is zero but the converse is unclear. When estimated parameters have zero variances, so that $\widehat{\mathbf{x}}_{T+h|T} = \mathsf{E}_T[\widehat{\mathbf{x}}_{T+h|T} \mid \widetilde{\mathbf{X}}_T^{T-r+1}, \mathbf{D}_{T+h}^{T+1}, \psi_T^e]$, then (iv) is zero and the converse holds *almost surely*. Finally, (v) is zero if and only if the world is non-stochastic.

Thus, the taxonomy includes elements of the main sources of forecast error, partitioning these by whether or not the corresponding expectation is zero. For there to be a gain from DMS, it must be obtained through estimation uncertainty (iv), possibly interacting with misspecification of deterministic or stochastic elements, (iia) and (iib). This is why the literature has shown that direct multi-step estimation is beneficial for forecasting essentially in two contexts: when the model is misspecified for the stochastic properties of the process (omitted unit roots or non-stationary regressors) or when deterministic properties alter and go unnoticed, as in the context of breaks, which may reinforce the previous type of misspecification via induced serial correlation of the residuals or long memory.

11. Conclusion

This paper has presented a review of the existing work on direct multi-step estimation for forecasting at varying horizons. We have shown that this strain of the literature has produced a vast amount of theoretical and empirical evidence favouring the use of this technique. Unfortunately, the diversity of approaches has made it difficult to draw a definite conclusion about its when's and why's. Here, we have shown that from the early contributions, the analyses have evolved towards either using DMS criteria for the design of forecasting models, or proper DMS estimation. In the light of our review, although the gain from using IMS or DMS varies with the horizon and the stochastic properties of the data, it is clear that the latter technique can be asymptotically more efficient than the former even if the model is well specified. This result is explained by the improvement in the variance of the multi-step estimator resulting from direct estimation. It thus appears that the misspecification of the error process in the case of DMS estimation is not so detrimental to the accuracy of the estimators. However, the limiting distributions reflect only partially the estimation properties of the methods. Indeed, in finite samples, the absence of bias can never be perfectly achieved and, hence, DMS can prove a successful technique for obtaining *actual* estimates and not only for reducing the multi-step variances – and indeed with respect to the latter IMS could prove more precise. There is little hope for success for DMS in finite samples when the data are stationary and the models are well specified. By contrast, when the models may be misspecified, *ex ante or ex post*, DMS provides accuracy gains, both asymptotically and in finite samples. As discussed in a general framework which allowed for a study of the various causes of forecast error, the main features that advocate DMS use are stochastic or deterministic non-stationarity. The literature showed that it could originate from breaks, unit roots, fractional integration or omitted regressors.

We can broadly separate the future research agenda into two categories. On the one hand, the existing trend on the analyses of models and circumstances will continue. The influences of breaks need be evaluated further, in particular using the link between occasional shocks and fractional cointegration. Co-breaking – linear combinations of variables which are insensitive to the breaks – would be valuable here. Non-linear estimation and breaks that occur after the forecast origin also need more study; so does the important issue of finite sample performance. Second, a fruitful strain revolves around model design. Recent work on the link between in-sample regressor collinearity and out-of-sample forecast performance seems an interesting route to pursue. In particular, the progress made in forecasting using factor analysis – when more variables than observations are available – point towards studying DMS properties since IMS is not an option in this context. Studies by Ing showed that it is possible to find algorithms that provide the most efficient *ex ante* multi-step forecasting technique for a stationary process. Further work on the features that advocate the use of DMS together with strategies for testing their presence would be valuable.

Acknowledgement

This paper reviews the literature available as of 2006. I am grateful to David Hendry, Bent Nielsen, Mike Clements and two anonymous referees for their helpful comments and suggestions.

Notes

1. That is, if and only if $\forall \eta > 0$,

$$\liminf_{T\to\infty}\left\{\min_{\boldsymbol{\theta}\in N_T^C(\eta)}[\overline{G}_{T,\overline{h}}(\boldsymbol{\theta}) - \overline{G}_{T,\overline{h}}(\widetilde{\boldsymbol{\theta}})]\right\} > 0$$

   where $\mathbf{N}_T(\eta)$ is a neighbourhood of $\widetilde{\boldsymbol{\theta}}$ of radius η such that its complement $\mathbf{N}_T^C(\eta)$ is a compact set of $\boldsymbol{\Theta}$.
2. We assume here that the econometric modeller does not *intentionally* misspecify her model. She therefore considers it to be the DGP.
3. We assume here that only these parameters are estimated, and that the functional forms are part of the models, so that we do not write $\widehat{\mathbf{g}}(\cdot)$ and $\widetilde{\mathbf{k}}(\cdot)$, as would happen in the 'non-parametric' models presented in Bhansali (2002).

References

Allen, P.G. and Fildes, R.A. (2001) Econometric forecasting strategies and techniques. In J.S. Armstrong (ed.), *Principles of Forecasting* (pp. 303–362). Boston, MA: Kluwer Academic.

Aron, J. and Muellbauer, J. (2002) Interest rate effects on output: evidence from a GDP forecasting model for South Africa. *IMF Staff Papers* 49: 185–213.

Bhansali, R.J. (1993) Order selection for linear time series models: a review. In T. Subba Rao (ed.), *Developments in Time Series Analysis* (pp. 50–56). London: Chapman and Hall.

Bhansali, R.J. (1996) Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics* 48: 577–602.

Bhansali, R.J. (1999) Parameter estimation and model selection for multistep prediction of time series: a review. In S. Gosh (ed.), *Asymptotics, Nonparametrics and Time Series* (pp. 201–225). New York: Marcel Dekker.

Bhansali, R.J. (2002) Multi-step forecasting. In M.P. Clements and D.F. Hendry (eds), *A Companion to Economic Forecasting* (pp. 206–221). Oxford: Blackwell.

Bjønstad, J. (1990) Predictive likelihood: a review. *Statistical Science* 5: 242–265.

Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis, Forecasting and Control* (2nd edn). San Francisco, CA: Holden–Day. First published 1970.

Chatfield, C. (1993) Calculating interval forecasts. *Journal of Business and Economic Statistics* 11(2): 121–135.

Chevillon, G. (2005) 'Weak' trend for estimation and forecasting at varying horizons in finite samples. *Oxford Economics Working Paper 210.*

Chevillon, G. (2006) Multi-step forecasting in unstable economies: robustness issues in the presence of location shifts. *Oxford Economics Working Paper 257*.

Chevillon, G. and Hendry, D.F. (2005) Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting* 21: 201–218.

Clark, T.E. and McCracken, M.W. (2005) Evaluating direct multi-step forecasts. *Econometric Reviews* 24: 369–404.

Clements, M.P. and Hendry, D.F. (1993) On the limitations of comparing mean squared forecast errors. *Journal of Forecasting* 12: 617–637.

Clements, M.P. and Hendry, D.F. (1996) Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* 58: 657–683.

Clements, M.P. and Hendry, D.F. (1998a) Forecasting economic processes. *International Journal of Forecasting* 14: 111–131.

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

Clements, M.P. and Hendry, D.F. (1999) *Forecasting Non-Stationary Economic Time Series*. Cambridge, MA: MIT Press.

Clements, M.P. and Hendry, D.F. (2004) Pooling of forecasts. *Econometrics Journal* 7: 1–33.

Clements, M.P. and Hendry, D.F. (2005) Guest editor's introduction: Information in economic forecasting. *Oxford Bulletin of Economics and Statistics* 67S: 713–754.

Cox, D.R. (1961) Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society B* 23: 414–422.

Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253–263.

Engle, R.F., Hendry, D.F. and Richard, J.F. (1983) Exogeneity. *Econometrica* 51: 277–304.

Ericsson, N.R. and Marquez, J. (1998) A framework for economic forecasting. *Econometrics Journal* 1: C228–C226.

Fildes, R.A. and Stekler, H.O. (2002) The state of macroeconomic forecasting. *Journal of Macroeconomics* 24: 435–468.

Findley, D.F. (1983) On the use of multiple models for multi-period forecasting. *Proceedings of Business and Economic Statistics, American Statistical Association*, 528–531.

Granger, C.W.J. (1969) Prediction with a generalized cost of error function. *Operations Research Quarterly* 20: 199–207.

Haavelmo, T. (1940) The inadequacy of testing dynamic theory by comparing theoretical solutions and observed cycles. *Econometrica* 8: 312–321.

Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica* 12 (Supplement): 1–115.

Hartley, M.J. (1972) Optimal simulation path estimators. *International Economic Review* 13: 711–727.

Harvey, A.C. (1993) *Time Series Models* (2nd edn). Hemel Hempstead: Harvester Wheatsheaf. First edition 1981.

Harvey, D., Leybourne, S. and Newbold, P. (1997) Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13: 281–291.

Harvey, D., Leybourne, S. and Newbold, P. (1998) Tests for forecast encompassing. *Journal of Business and Economic Statistics* 16: 254–259.

Hausman, J. (1978) Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Haywood, J. and Tunnicliffe-Wilson, G. (1997) Fitting time series model by minimizing multistep-ahead errors: a frequency domain approach. *Journal of the Royal Statistical Society B* 59: 237–254.

Haywood, J. and Tunnicliffe-Wilson, G. (2004) A test for improved multi-step forecasting. *Mathematics and Statistics Research Reports*, Victoria University of Wellington.

Hendry, D.F. (2000) A general forecast-error taxonomy. Econometric Society World Congress 2000, Contributed Papers 0608, Econometric Society.

Hendry, D.F. and Hubrich, K. (2005) Forecasting aggregates by disaggregates. Mimeo, Nuffield College, Oxford.

Hoque, A., Magnus, J.R. and Pesaran, B. (1988) The exact multiperiod mean square forecast error of the first-order autoregressive model. *Journal of Econometrics* 39: 327–346.

Hurvich, C.M. (2002) Multistep forecasting of long memory series using fractional exponential models. *International Journal of Forecasting* 18: 167–179.

Ing, C.-K. (2003) Multistep prediction in autoregressive processes. *Econometric Theory* 19: 254–279.

Ing, C.-K. (2004) Selecting optimal multistep predictors for autoregressive process of unknown order. *Annals of Statistics* 32: 693–722.

Johnston, H.N. (1974) A note on the estimation and prediction inefficiency of 'dynamic' estimators. *International Economic Review* 15: 251–255.

Johnston, H.N., Klein, L. and Shinjo, K. (1974) Estimation and prediction in dynamic econometric models. In W. Sellekaerts (ed.), *Essays in Honor of Jan Tinbergen*. London: Macmillan.

Kabaila, P.V. (1981) Estimation based on one step ahead prediction versus estimation based on multi-step ahead prediction. *Stochastics* 6: 43–55.

Kang, I.-B. (2003) Multiperiod forecasting using different models for different horizons: an application to U.S. economic time-series data. *International Journal of Forecasting* 19: 387–400.

Klein, L.R. (1971) *An Essay on the Theory of Economic Prediction*. Chicago, IL: Markham.

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

Liu, S.I. (1996) Model selection for multiperiod forecasts. *Biometrika* 83(4): 861–873.

Madrikakis, S.E. (1982) The accuracy of time series methods: the results from a forecasting competition. *Journal of Forecasting* 1: 111–153.

Mann, H.B. and Wald, A. (1943) On the statistical treatment of linear stochastic difference equations. *Econometrica* 11: 173–220.

Marcellino, M., Stock, J.H. and Watson, M. (2006) A comparison of direct and iterated multistep AR methods for forecasting microeconomic time series. *Journal of Econometrics* 135: 499–526.

Peña, D. (1994) Discussion: Second-generation time-series model, a comment. *Journal of Forecasting* 13: 133–140.

Rissanen, J. (1986) Order estimation by accumulated prediction errors. *Journal of Applied Probability* 6: 67–76.

Sargan, J.D. (1982) On Monte Carlo estimates of moments that are infinite. *Advances in Econometrics* 1: 267–299.

Schorfheide, F. (2005) VAR forecasting under misspecification. *Journal of Econometrics* 128: 99–136.

Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8: 147–164.

Stoica, P. and Nehorai, A. (1989) On multi-step prediction errors methods for time series models. *Journal of Forecasting* 13: 109–131.

Stoica, P. and Soderstrom, T. (1984) Uniqueness of estimated k-step prediction models of ARMA processes. *Systems and Control Letters* 4: 325–331.

Tiao, G.C. and Tsay, R.S. (1994) Some advances in non-linear and adaptive modelling in time-series analysis. *Journal of Forecasting* 13: 109–131.

Tiao, G.C. and Xu, D. (1993) Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika* 80: 623–641.

Tsay, R.S. (1993) Comment: Adaptive forecasting. *Journal of Business and Economic Statistics* 11 (2): 140–142.

Weiss, A.A. (1991) Multi-step estimation and forecasting in dynamic models. *Journal of Econometrics* 48: 135–149.

Weiss, A.A. (1996) Estimating time series models using the relevant cost function. *Journal of Applied Econometrics* 11: 539–560.

Weiss, A.A. and Andersen, A.P. (1984) Estimating time series models using the relevant forecast evaluation criterion. *Journal of the Royal Statistical Society A* 147: 484–487.