"THE IMPACT OF EMPIRICAL ACCURACY STUDIES ON TIME SERIES ANALYSIS AND FORECASTING"

by

Robert FILDES*
and
Spyros MAKRIDAKIS**

93/29/TM

* Professor, Department of Operational Research and Operations Management, at Lancaster University, Lancaster LA1 4YX, U.K.

** Research Professor of Decision Sciences and Information Systems, at INSEAD, Boulevard de Constance, 77305 Fontainebleau, France.

Printed at INSEAD, Fontainebleau, France

THE IMPACT OF EMPIRICAL ACCURACY STUDIES ON TIME SERIES ANALYSIS AND FORECASTING

Robert Fildes
Department of Operational Research
and Operations Management
Lancaster University
Lancaster LA1 4YX
Tel: (0)524 65201
Fax: (0)524 844885
e-mail: ora004@uk.ac.lancaster.cent1

Spyros Makridakis
INSEAD
Boulevard de Constance
77305 Fontainebleau
France
Tel: (1) 60 72 42 36
Fax: (1) 60 72 42 42

ABSTRACT

Social scientists envy the objectivity, controlled experimentation and replicability of hard sciences, a lack of which, they claim, hampers their ability to advance their disciplines and make them more useful and relevant to real life applications. This paper examines a specific area of social science, time series forecasting, which, through empirical studies using real-life data, allows for objectivity and replicability and offers the possibility of controlled experimentation. Yet its findings are ignored and its conclusions to advance the field of forecasting are disputed. The paper describes what has been learnt from forecasting competitions and compares the results with expectations based on statistical theory. It demonstrates that considerable anomalies exist which have been neglected by academic statisticians who have focussed their attention on topics/directions of little practical value, and no relevance for real-life applications The paper concludes with a challenge to theoretical statisticians and empirical researchers alike: working together they can learn from each other and advance their field to better serve the business and economic communities and make their area more useful and relevant to policy and decision makers eager to use more accurate predictions. Equally important, forecasting competitions can provide them with objectivity, replicability and controlled experimentation that can direct progress in their discipline.

Key Words: Forecasting, Time Series Models, Philosophy of Science, Replication, Citation Analysis, Statistical Paradigm

THE IMPACT OF EMPIRICAL ACCURACY STUDIES ON TIME SERIES ANALYSIS AND FORECASTING

Robert Fildes
Department of Operational Research
and Operations Management
Lancaster University
Lancaster LA1 4YX
Tel: (0)524 65201
Fax: (0)524 844885

Spyros Makridakis
INSEAD
Blvd. de Constance
Fontainbleau
France
Tel: (1) 60 72 42 36
Fax: (1) 60 72 42 42

". . . I shall be surprised if it (the Makridakis/Hibon (1979) study) has any real impact on the statistical audience, towards which it appears to be so bravely aimed ", Oliver Anderson, (1979)

Revised April 1993

1. INTRODUCTION

Social scientists in general and those working in the fields of economics and management in particular often view hard sciences as providing them with a model to imitate. Most of all they are impressed by their apparent high predictability as well as their ability to establish universal laws that allow highly accurate predictions. Furthermore, they are envious of controlled experimentation in fields like physics and chemistry and the replicability and objectivity that characterize these disciplines. For instance, it took exactly one year to publish a paper in the March 1990 issue of *Nature* that disproved beyond reasonable doubt claims advanced by two chemists (Stanley Pons and Martin Fleischmann) on March 23, 1989 that they had harnessed fusion energy in a test tube of water at room temperatures (Close, 1990). In the social sciences similar refutation would have taken years to advance, even longer to publish and probably ended in argument and counter argument with no definite conclusions.

Can social scientists expect that their fields will achieve the universality and objectivity of hard sciences and become capable of controlled experimentation and replicability of their findings? These are important questions that need to be addressed within the fields of economics and management (and those of other social sciences) in their attempt to become useful and relevant for decision and policy makers. This paper examines time series forecasting as a discipline that offers a methodology for research in the field of social sciences and attempts to answer these questions within the framework of empirical studies aimed at studying the accuracy of various forecasting methods. The heart of our argument is that in time series forecasting, experimentation and replication are possible. However capitalizing on their benefits can only be achieved if the implications of successful empirical research are acknowledged and utilized by theoreticians. Here, we argue, is where the statistical community has failed and where a major effort is needed to learn from empirical research in order to advance the field of time series forecasting.

The Structure of Scientific Revolution

Karl Popper (1972) and his followers have popularized the notion of "falsification" of existing scientific theories, arguing that no theory can ever be considered as a 'universal truth' since it is only a matter of time before it will be rejected by becoming inconsistent with empirical data. Kuhn (1962), although agreeing with Popper, went an important step further by showing that scientists tend to continue working within their maintained paradigm even after empirical evidence has accumulated suggesting considerable anomalies between the deductions derived from the espoused theory and what

is observed. He therefore postulated that science does not evolve slowly under the impact of new evidence but instead such evidence is ignored for as long as possible, in particular if it is in conflict with the basic tenets of the accepted paradigm. Science, he argues, progresses in revolutionary steps only when the anomalies become so strong that no *ad hoc* modifications in the accepted theory can explain the empirical observations. But even in such cases, fundamental changes do not take root until the "old guard" of researchers is replaced by a new generation of scientists whose minds are open to the new theories (and whose careers may prosper by following a new line of research).

It is important to realize that Kuhn based his observations on the hard sciences, where replicability and objectivity are the norms, and not the social sciences where they cannot usually be assured because controlled experimentation is neither easy nor practical. Although concerns about scientific progress in the social sciences are well beyond the range of this paper we examine progress in the field of time series forecasting as an example of the conflict between empirical evidence and the accepted statistical paradigm within the area. The next section describes the core assumptions underlying time series analysis and forecasting. The related empirical research, carried out over the last two decades, is discussed in Section 3 where the results of various studies of comparative accuracy are evaluated and contrasted with what might be expected if the accepted statistical paradigm held true. These studies have provoked a number of criticisms which we consider in Section 4 in order to assess their validity. Section 5 attempts to reconcile the empirical evidence with the expectations from statistical theory, and offers some suggestions as to the statistical advances desirable if the gap between the two is to be bridged. In concluding it is argued that specific hypotheses can (and have) been proposed in time series forecasting and these can (and have) been tested in replicable and objective 'experiments'. The results (supplemented, perhaps, by more finely focused research) should be used to guide future developments in the field to make it more useful and relevant to those attempting to interpret time series data and forecast on the one hand and those wishing to develop statistical theory on the other.

2. THE PARADIGM OF TIME SERIES FORECASTING

Although no history of time series has been written (see Makridakis, 1976), an examination of the early books in the area such as Wold (1938), and Grenander and Rosenblatt (1957) or Hannan (1960) shows the prevailing modelling tradition to have been that of linear stationary models. The publication of Box and Jenkins' (1970) book popularized time series methods by providing a practical approach for modelling linear non-stationary processes that included autoregressive and/or moving

average components; their approach also extended to include multiple time series situations. The methodology they proposed was that the variable of interest was first transformed to stationarity (this could then permit using the available theory which was only applicable for stationary data), an appropriate ARIMA model then had to be specified using the transformed data, its parameters estimated by fitting such a model to available historical data, and its validity verified by diagnostic checking of the residuals, in particular their independence. The accepted statistical paradigm was that the correctly specified model that best fitted the historical data would also be the optimum model for forecasting purposes.[1] While it has been argued that *Time Series Analysis, Forecasting and Control* contained nothing that was new to the time series statistician (Priestley, 1973) it nevertheless attracted considerable attention by researchers within and outside the statistical field. Its strength was that it dealt with practical considerations with a strong theoretical framework which when applied could effectively model a very wide range of time series (Box and Jenkins, 1973) or for practical purposes all such series (Jenkins, 1974). The availability of computers in the 1970s further contributed a significant growth in time series methods with the ARIMA formulation used as the starting point for further theoretical developments and applications.

The focus of statistical research into time series over the past two decades can be established unequivocally by examining key journals. Using the Science Citation Index as a guide the *Journal of the American Statistical Association* (JASA) and that of the Royal Statistical Society have had the highest impact (JASA has published the largest number of articles, approximately 150 per year, some 40% more than the three journals of the RSS combined). By examining each time series article published in the years 1971-1991 a number of key words were established as useful in describing those articles that might be relevant to the study of univariate time series forecasting methods, suitable for tackling business and economic problems. The reliability of the classification was established by a research assistant going through the Journals and keywording each article. The classification scheme used had the dimensions shown in Table 1 (with further details in Table 4).

[1] There has been a long tradition of automatically fitting time series models to data although there has been considerable argument as to the appopriate criterion (See de Gooijer et al. (1985) for a survey). Box and Jenkins (1970) develop Occam's razor into the principle of parsimony but rely on the data alone for the evidence. Later writings by them talk more of matching the theoretical structure (be it economic or physical) to the characteristics of the model. However all these approaches are based on using within-sample information only.

KEYWORD	% Total: 1971-1980	% Total: 1981-1991	Total Number
Model			
- ARIMA	45%	55%	97
- State Space	5%	10%	20
- Non-linear	3%	5%	11
- Trend Curves (and time series splines)	2%	7%	13
Approach			
- Theory	39%	21%	
- Theory with illustration	15%	48%	
- Empirical Study	46%	31%	
TOTAL REFERENCES	**64**	**177**	**241**

Table 1 Models and Approach adopted in Time Series Research : 1971 - 1991

N.B. These figures are subject to revision.

An examination of the above table confirms that the ARIMA paradigm has remained dominant although in recent years State Space modelling has been given increasing attention. Moreover, the majority of articles have been concerned with theoretical contribution rather than addressing practical problems.

In distinct contrast to the statistical model building approach, management scientists and operational researchers, as exemplified by Brown, proposed a series of *ad hoc* forecasting methods, designed to be suitable for successful implementation. Their statistical foundations had been set out in Brown's book, *Smoothing, Forecasting and Prediction*, (1962) and centred around various generalised exponential smoothing models. Despite the early interest of statisticians such as Cox (1961) in its robustness, exponential smoothing was seen (incorrectly) as nothing more than a special case of an ARIMA model, and its apparent success in being widely adopted in business applications during the 1960s and 1970s did not encourage much further research into its properties (but see Harrison, 1967). This is an early illustration where empirical information (no matter its crudeness and lack of sophistication) about the superiority of exponential smoothing models for improving forecasting accuracy in business applications was completely ignored by practically all theoretical statisticians and academic forecasters.

3. FORECASTING ACCURACY STUDIES

Comparative reviews of the forecasting accuracy of alternative forecasting methods established themselves early as an adjunct to the theoretical developments of economic and statistical models. As early as 1956 Ferber (1956) and Shupack (1962) showed concern that 'correctly specified', well-fitted models often under performed when tested outside the sample data. However until the early 1970s there was limited empirical estimation of sophisticated forecasting models as computer power, memory and availability were still restricted. Thus, there were few opportunities to verify the accuracy of predictions derived from the accepted paradigm that the best fitting model also produced the best forecasts, (the assumption of constancy in our terminology). It was not until the late 1960s that forecasting accuracy studies started focusing on a wide range of comparisons that covered several methods and used real data in order to determine the accuracy of such methods. The greater availability, lower cost and improved speed and memory of computers made such empirical studies easier permitting an enlarged number of series and methods to be compared.

A Brief History of Forecasting Accuracy Studies

Reid (1969), and Newbold and Granger (1974) using 106 time series concentrated on the accuracy of three univariate forecasting methods: the Holt-Winters variant of exponential smoothing with trend and multiplicative seasonal, stepwise autoregressions on first differences of the data with automatic selection of lag length, and the Box-Jenkins methodology to ARIMA modelling. They also considered in depth various forms of combining forecasts from these three methods. With the exception of some studies comparing univariate time series models to econometric models such as that by Prothero and Wallis (1976) no further ambitious accuracy comparisons were undertaken until Makridakis and Hibon (1979) who, like Newbold and Granger before them, used the hospitality of the Royal Statistical Society to present the findings of their empirical study.

The Makridakis/Hibon study was based on 111 series. Where it differed from the Newbold/Granger study was in the number of methods included: some 13 core methods[2] in total including a number of variations of exponential smoothing, adaptive smoothing and naive benchmarks based on the random walk. While the Newbold/Granger study offered findings that confirmed (albeit weakly) the strongly held opinions of the leading time series statisticians of the day, the results of the Makridakis/Hibon study were more openly in conflict. Such conflict was not received well by the statisticians of the time who tried to find fault with the study and its findings. Further, commentators made a number of

---

[2] We have counted each method that has both a seasonal and non-seasonal variant as a single method.

substantive suggestions (see below) on how the design of comparative studies could be improved. This led Makridakis to continue this line of research which resulted in the so called M-Competition (Makridakis *et al*, 1982).

In the M-Competition, the number of series considered was increased to 1001 and the number of core methods was also augmented to 15. In addition more accuracy measures were considered and the series were segmented into various subdivisions in search of an explanation as to the circumstances in which one method outperformed the remainder. Like the earlier major studies a commentary was published (Armstrong and Lusk, 1983) that identified weaknesses, clarified results and suggested paths for further research. Newbold (1983) even hoped that it was 'the forecasting competition to end all forecasting competitions'. However further research has been carried out with the intent of extending the original findings, clarifying some aspects of the methodology (Fildes, 1992, Armstrong and Collopy, 1992), extending the range of univariate methods (Gardner and McKenzie, 1985), including multivariate comparisons (Kling and Bessler, 1985) and, following Newbold (1983), examining comparative performance under conditions more closely matching those experienced by practising forecasters (the M2- Competition, Makridakis *et al*, 1993).

Empirical Results and Statistical Anomalies from Forecasting Competitions

Starting with Reid (1969) and the two Makridakis studies many additional empirical comparisons have been conducted that have identified various "anomalies" between the predictions deriving from statistical theory and the empirical evidence. The critical assumptions of time series modelling and the resulting "anomalies" are described next.

Stationarity: Stationarity is an assumption made in most, if not all, early statistical modelling approaches. Box and Jenkins (1970) argue that the data are best transformed to stationarity by differencing. Quite early in the research into empirical performance, Pierce (1977) argued that the differencing transform was inadequate and that a simple linear time trend was often more appropriate. Various studies have identified this issue as being of both statistical and economic importance (see Nelson and Plosser, 1982) with recent statistical work pointing to the difficulties in distinguishing between the two specifications through a model of the form:

$$Y_t = \alpha + \beta t + \delta Y_{t-1} + e_t$$

While Dickey and Fuller have proposed certain tests of this and similar models they have been shown to have low power against similar alternatives (DeJong *et al.*, 1991). Fortunately the results of the

various accuracy studies have cast some light on this with Makridakis *et al.* (1982) demonstrating considerable success with Parzen's method (Parzen, 1982, Newton and Parzen, 1984) which estimates from the data a prefilter consisting of a long memory autoregression. For example, as measured by MAPE, Parzen's approach outperforms the ARIMA formulation 52% of the time with an average improvement of about 15% depending on lead time. Meese and Geweke's (1984) results added support to this with differencing performing worse than linear detrending and about equivalently to an estimated prefilter of lagged values.

The relative success of simple exponential smoothing (with no trend) compared with Holt's method which extrapolates a linear trend estimate (Makridakis and Hibon (1979), Makridakis *et al.*(1982)) and the subsequent development of Gardner and McKenzie's (1985) damped trend all underline the importance of trend and the need to identify it and extrapolate it correctly. In summary, important differences, in particular for longer-term forecasting, in accuracy can result depending on how the various univariate approaches deal with trend in the data, while conventional within-sample testing is unlikely to reveal the best approach.

Logarithmic transformations and other Non-Linearities: A model may need transforming because the data generation process is non-stationary, or because the estimation procedure employed requires a stationary error term. The pursuit of the appropriate transformation and its efficacy has been controversial, see for example Chatfield and Prothero (1973) where various researchers argued over the merits of transforming a data series to different powers (from .34 to 0, equivalent to a log transform) when only a single series was under analysis. Of the major empirical studies, Makridakis and Hibon (1979), Makridakis *et al.* (1982) and Meese and Geweke (1984) considered the problem. The results were highly disturbing. In Meese and Geweke 48% of the series apparently needed a log transform but when out-of-sample criteria were used for evaluation there was disagreement in about 45% of the series with the in-sample recommendation, i.e. the information from the in-sample analysis was close to worthless. In the Makridakis/Hibon study transformations produce worse forecasts than no transformations while in the M-Competition no difference was found. In summary, transformations to achieve a constant variance have been shown to have no effect on out-of-sample forecasting accuracy yet they are indispensable according to statistical theory.

Sophisticated vs. Simple Methods: Because the Box-Jenkins methodology permits the user to select an appropriate theoretical model to best match the characteristics of the data,[3] it could be reasonably expected that the theoretical correctness and the additional flexibility would lead to improved accuracy. However, this has not happened. Newbold and Granger (1974) found only limited improvements in accuracy when they compared ARIMA models to the automatic Holt-Winters version of exponential smoothing. In contrast, both Makridakis and Hibon (1979) and Makridakis *et al.* (1982) found support for the view that simple methods, by which was meant mechanical methods such as exponential smoothing or even the benchmark methods of a random walk, or a seasonal variant, outperformed more complex specifications such as the ARIMA. Additional empirical research has given further support to the view that simple time series models do at least as well as statistically sophisticated ones: Schnaars (1986), Koehler and Murphee (1988), Huss (1985), Geurt and Kelly (1986), Watson *et al.* (1987), Collopy and Armstrong (1992), Fildes (1983), Makridakis et al. (1993). Many of these studies use different data sets and a variety of methods.

Despite, or perhaps bacause of, the lack of success of sophisticated methods researchers have chosen to widen the class of models they considered. An example is the state dependent class of model proposed by Priestley (1988). However empirical evidence on their performance is weak and as DeGooijer and Kumar (1992) point out, unconvincing. Thus, no improvements in out-of-sample accuracy have been proven from adopting more sophisticated models, despite persuasive statistical within-sample evidence.

General and Special Cases: It can be shown that many variants of exponential smoothing are special cases of ARIMA models (Cogger, 1974, Gardner and McKenzie, 1985). It was therefore argued that it was not possible for the latter to outperform the former, performance could at best be equal.[4] Starting with Makridakis and Hibon (1979) evidence has accumulated that ARIMA based forecasting even when applied strictly according to the authors' original intentions is outperformed by exponential smoothing, This result was supported by Makridakis *et al.* (1982) and (1993). However, even when it is true (not the case for many non-seasonal and additive seasonal models) that exponential smoothing is a special case of ARIMA models there is no way to theoretically assure that the out-of-sample accuracy of ARIMA models will exceed those based on exponential smoothing. The only claim that can be made is that the model fitting errors of ARIMA models will be as small as those of exponential smoothing methods which empirical studies have shown to be true.

---

[3] Automatic variants of ARIMA modelling and state space approaches match the data characteristics objectively by employing various model selection criteria to identify the appropriate model structure.

[4] While not explicitly claimed in Jenkins' (1974) comments on Newbold and Granger (1974) that is the tenor of his remarks.

Forecasting Accuracy and Prediction Intervals: All forecasts are most likely (or expected) values and are associated with an interval that expresses uncertainty. Empirical studies (Makridakis and Winkler, 1989, Makridakis *et al.*, 1987) have shown, however, that actual forecasts fall outside the theoretically constructed confidence intervals more often than postulated by the theory, whatever model is used to describe the data. See Chatfield (1993) for a survey of this question. For example, in Makridakis *et al.* (1987) it was shown that 17% of the forecasts fell outside the 95% confidence interval for lead 1 rising to 26% for lead 6.

The Effect of Sample Size: According to statistical theory the size of the prediction interval, directly related to the standard error, ought to decrease according to the square root of the sample size used in model estimation. Empirically, however, this result has not been confirmed in Makridakis and Hibon (1979), Makridakis *et al.* (1982), and Lusk and Neves (1984). Such a result contradicts a major tenet of statistical theory and raises some fundamental questions about the optimal sample size about real-life applications, the obvious conclusion being that there is no need to search for more data (an expensive and time consuming task) when less can do as well.

Combining: Starting with Newbold and Granger (1974) and supported by Makridakis and Hibon (1979), the M-Competition (1982), the M2-Competition (1993) and a large number of other empirical studies both inside and outside the field of forecasting (see Clemen 1989 for a review and annotated bibliography) the conclusion has been maintained that combining more than one forecasting method (at least one of which will naturally be suboptimal for a given data set) results in more accurate out-of-sample forecasts. In addition, to add insult to injury, empirical research has found that simply averaging the forecasts of the various methods is as accurate as combining them according to some optimizing procedures that minimize the variance and/or covariance of the methods being combined.

Fit vs Forecast Performance: If superior model fitting performance results in better out-of-sample forecasting accuracy then there should be a close correlation between the two. However, this is not the case; Makridakis (1986) and Makridakis and Winkler (1989) have found that such correlations are around 0.2 (i.e., only 4% of the post-sample accuracy is explained) for the first three forecasting horizons, which then drop towards 0.1 by period five and zero by period 12. Similar conclusions (based on the M-Competition data) have been reached by Pant and Starbuck (1990), although using MAPE as a measure leads to some improvement. If a close relationship between model fit and out-of-sample forecasts does not exist it is hard to argue that model selection can be

based on minimizing model fitting errors. Moreover, there is no reason to guarantee that a certain method will perform better than others because it better explains the past, or because its model fitting errors are smaller. In effect, one can argue that the use of model fitting criteria might turn out to be more sophisticated variants of the case of fitting an n-1 degree polynomial to a set of n data points to achieve zero model fitting errors. Forecasting errors will not be zero in practically any application.

To better understand the issues involved (including the anomalies described above), we concentrate on four of the major large scale empirical studies, Newbold and Granger (1974), Makridakis and Hibon (1979), Makridakis *et al* (1982) and Meese and Geweke (1984). The objectives of these studies are listed, their 'experimental design' described and the objections voiced by the critics are discussed in the next section.

The Objectives and Experimental Design of Forecasting Accuracy Studies

Newbold and Granger (1974) set themselves some modest objectives. Having noted that certain forecasting methods were 'fully automatic' they aimed 'to assess the potential loss in terms of forecasting accuracy' in using such methods compared with those that required subjective inputs. More generally their study was designed to 'assess the relative performance of some [univariate forecasting] methods on real data'. These two ideas were combined by the authors when they said, in responding to the commentators at their paper presentation to the RSS, that they did not expect 'automatic procedures to perform as well as those requiring subjective inputs (based on the data characteristics alone) such as Box-Jenkins; another objective was to get an impression as to how much might be sacrificed in terms of forecasting accuracy by employing them'.

When summarizing additional accuracy studies in addition to that of Newbold and Granger, Makridakis and Hibon (1979) observed conflicting evidence on the supposed better performance of the Box-Jenkins approach. Their aim then was to reconcile these apparent disagreements (which were highlighted in the earlier section) and, in addition, figure out the reduction in forecasting accuracy that was associated with the usage of simpler methods that, in their experience, were employed to a much greater extent in business firms and the military than ARIMA modeling. A consensus view emerged from the RSS discussion of the Makridakis and Hibon paper that an explanation for observed differences should be sought in the characteristics of the time series (explored in the Makridakis and Hibon study somewhat unsuccessfully, at least as judged by the RSS's commentators). Further, such information was to be used to help "forecasting users [make] rational choices for their situations".

Makridakis *et al.*(1982) picked up on this theme with an introductory statement that "what is important, therefore, is not to look for 'winners' or 'losers', but rather to understand how various forecasting approaches and methods differ from each other and how information can be provided so that forecasting users can be able to make rational choices for their situations." They went on to liken their approach to the testing of consumer products by measuring and comparing their various features.

Like the earlier authors, Meese and Geweke (1984) focus on comparative accuracy, but as Table 2 below shows, they went further along the road of specifying explicit hypotheses as to those factors that could affect the outcomes: (a) data transformations such as logs, detrending and differencing, (b) data periodicity, (c) the forecast horizon, (d) the criterion used to measure accuracy, (e) the loss function employed in parameter estimation and (f) the seasonal adjustment procedure.

The Experimental Designs

Table 2 below summarizes the key experimental characteristics of the four studies we have examined. It shows that such studies have covered a wide variety of methods, used a large number of series, and utilize practically all loss functions suggested in theory and available in practice.

Experimental Characteristics	Research Studies			
	Newbold/Granger	Makridakis/Hibon	Meese/Geweke	Makridakis *et al.*
Methods	3 methods + combining (ARIMA, exponential smoothing, & Autoregregression)	12 methods + seasonal variants (ARIMA, exponential smoothing)	Autoregressive Models - Various model selection criteria, e.g. AIC, BIC	13 Methods + seasonal variants + combining (ARIMA, Bayesian, Exponential Smoothing)
- Estimation	Least Squares	Least Squares	Least Squares, MAD	Least Squares
Data	106 Series: 80 monthly, 20 quarterly; mixture of micro and macro, some seasonal; data trended	111 series; 80% monthly, 12% quarterly, 8% annual: 1/3 macro, 2/3 seasonal. Data did not consistently trend	150 series, all macro, 50 quarterly, 100 monthly, 1/3 seasonally adjusted. 40% financial, 60% real, 1/3 non-US	1001 series + Sub-sample of 101: 302 micro data, 236 industry, 319, macro and 144 demographic, with 181 annual, 203 quarterly and 617 monthly
Lead Times	1-8	Leads 1-6, 9, 12	Leads 1,6,12	Leads 1-18
Loss Functions	Distribution of Relative Average Squared Errors (Lead 1), % better for all leads	Fit statistics: MAPE, Theil's U; % better	% better measured by Relative MSE, MAPE and relative error: no fit or forecast statistics given	MSE, MAPE, Median APE, % better, Average Rankings
Transforms	None	Log and square root in ARIMA models	Logs to maximize likelihood; various prefilters of data	Log transforms for ARIMA models, none for other methods

Table 2 The Experimental Characteristics of Accuracy Studies

In summary, the objectives of the empirical studies have remained consistent over time, with an increasing emphasis on explaining observed differences to help the forecaster select the most accurate forecasting method for the specific situation at hand, while their experimental characteristics are such as to cover all major theoretical and practical aspects of interests.

The Objections to Forecasting Accuracy Studies

Of the four competitions considered, the three that have encouraged commentary have provoked a number of objections ranging from the competence (or better, incompetence) of the researchers, to detailed statistical comments. We will limit our remarks to those that have been concerned with the broader issues discussed in the preceding sections, arguing point by point that these objections are not sufficient, either individually or collectively, to dismiss the conclusions drawn as irrelevant or caused by faulty experimental design.

Lack of clear objectives: In the various discussions of forecasting accuracy studies, a number of commentators have questioned the value of such comparisons: Priestley, commenting first in Newbold and Granger (1974) and then more explicitly on Makridakis and Hibon (1979), states "we should resist the temptation to read too much into the results of the analyses". He points out that the "results of the (Newbold and Granger) forecasting study tell us rather more about the models of the series analyzed that the relative merits of different types of forecasting methods". Newbold (1983) in a partial recantation of his earlier work doubts the value of such 'horse races' and goes on to note that the results that are reported are necessarily **aggregate**, so that "the forecaster, faced with a specific problem learns little about how such a problem might be attacked".

In commenting on the Newbold and Granger study, Reid (1974) highlighted the problem that without a well-specified population, selection of the time series to analyze cannot be based on sampling theory, but must, instead, be done so that the series are as representative as possible of the problem the researchers choose to address. To help in the interpretation of the results, Reid argued that the characteristics of the series chosen should be fully described. Durbin (1979) and Newbold (1983) took up this same criticism of the subsequent Makridakis studies, with Newbold stating that no inference on relative forecasting performance was possible.

Many (if not all) field experiments suffer from this same limitation. It is overcome by the experimenter increasing the sample size and the diversity from where the sample of time series are drawn and by assuming that the non-random components in the experiment are of little apparent importance to the possible outcome. As Johnstone (1989) argues, the statistical testing of hypotheses is not dependent on randomness in the sample, only lack of systematic bias. Inexplicable findings should lead to a revised definition of the population under study and the results expected from the treatment being tested. Reid's plea for a stratified random selection of series was responded to, in part, by Makridakis *et al.* in that relative accuracy was examined in sub-sets of the data: seasonal/non-

seasonal, micro/ macro etc. However quantitative measures such as the number and position of outliers, level of randomness (included in Makridakis and Hibon), trend/cycle component relative to noise could have remained important factors influencing the results in some systematic way.

Aggregation over lead times of error statistics: In the two Makridakis studies, only one fixed lead time error is calculated for each series which, Jenkins (1982) argues, limits their relevance. Errors at different lead times (for the same series) are typically correlated and therefore averaging errors over lead times (although of interest as a summary statistic in its own right) does not effectively increase the 'degrees of freedom' of the summary statistic. In an evaluation of summary error statistics and their reliability, Armstrong and Collopy (1992) have also shown that use of a cumulative error measure (over a number of lead times) does not lead to improved reliability compared to using the fixed lead time alternative. Newbold and Granger, Meese and Geweke and Makridakis *et al.* all avoid this pitfall by using fixed lead time statistics. The first two studies also accumulate (fixed lead time) errors for a given series across time. However for leads greater than one, the errors are still correlated. Thus, Jenkins (1982) claim that the one step ahead residuals contain all the information necessary in an evaluation is not correct unless the estimated model is appropriate for the out-of-sample range. This is rarely the case which makes observed differentials in comparative performance across lead time potentially suggestive of alternative modelling/estimating procedures.

Use of automatic methods of forecasting: Starting with Jenkins (1974, 1982) he and many other commentators have concerned themselves with the appropriateness of automatic methods. But as Bramson (1979) and Chatfield (1986) have made clear, automatic methods are not an indulgence but a necessity when dealing with the large inventory/product systems that many organizations have implemented. In addition, there is no evidence on the benefits of personalized model building (excluding domain knowledge) compared to automatic procedures. For example Hill and Fildes (1983), Libert (1983), Texter and Ord (1987), and Makridakis et al. (1993) all found no substantive improvements from personalized model building. At the very least an effective automatic benchmark would provide monitoring information for subjective adjustments based on expert knowledge (Fildes and Beard, 1992). It should not be presumed, as some of the commentators believe, that the best modelling can be done through an ARIMA model building approach, although Wu *et al* (1991) have shown how this can sometimes be achieved effectively.

Aggregation over time series of error statistics: Jenkins (1982) also argues against aggregating error statistics over time series, pointing out that the errors for different series may have a different population mean. However this criticism is without merit if we once concede that the series are in some sense representative of a population of time series. Alternatively, relative error measures can be used or the errors may be standardized. In the concrete case of choosing a method for a particular production or inventory control scheme, the series analyzed can be chosen to conform to a desired sampling scheme and appropriate error statistics calculated in ways that have managerial significance, e.g. safety stocks and their costs.

Use of a single time origin to construct the forecasts: The error statistics calculated are in the two Makridakis studies a summary of a cross section of forecast errors for different lead times. Newbold and Granger, and Meese and Geweke aggregate (fixed lead time) errors across time for each series and then publish summary statistics across all series. Although it is possible that the forecasts and their accuracy are not independent (e.g., if they all refer to a period of an economic boom) the problem is minimal when the series are selected in such a way as to end in different time periods and only one set of forecasts is done for each series (this was the case in the Newbold and Granger, Makridakis and Hibon and the M-Competition studies).

Failure to match estimation methods to error statistics: A number of commentators, e.g. Priestley (1974, 1979), Zellner (1986) have criticized the studies for estimating the parameters in a model using least squares, adopting the conditional mean as the point forecast and subsequently evaluating forecast accuracy using other measures than mean squared error, e.g. MAPE. Whilst such a procedure might have had significant effects on the results, other researchers have explored this issue and found little if any improvement when optimal estimators with better theoretical performance characteristics are used, and performance is evaluated out-of-sample using the matching evaluation criterion (Meese and Geweke, 1984, Makridakis and Hibon, 1991, see also Fildes and Makridakis, 1988).

A Refutation: Objectivity, Replicability and Methodological Adequacy

The conclusions from accuracy studies could be undermined if the researchers involved were seen to have any interest in introducing biases into the studies, either intended or unintended. On the contrary, for the studies considered here, those involved had a strong motive in showing that forecasting accuracy improved with increasingly sophisticated methods. (Speaking from personal

experience, both authors were surprised by some of the results of the M-Competition but felt that it was not appropriate to understate their implications as was repeatedly suggested to them. 'Are you sure you want to publish this stuff, it will destroy the field of forecasting?' was a common comment). Most importantly, as far as the M-Competition was concerned, every effort was made to achieve as high a degree of objectivity as possible. This included finding participants knowledgeable enough to carry out each method expertly and attempting to ensure that the procedures were well documented and available for later use[5].

An additional check on possible biases is through other researchers having the potential to first, replicate and then extend the studies. Replication has proved highly problematic in certain areas of empirical economics (Dewald, 1986) with failure arising from a variety of causes such as lack of explicit documentation, calculation error or data errors. When replication has been achieved the methodological approach that was chosen has turned out to be sensitive to further testing thereby undermining the results (Kramer, 1985). This problem has been responded to in the accuracy studies as follows. A sub-set of the Newbold and Granger (1974) study was further analyzed by Chatfield (1978). The M-Competition authors went further by making the data and forecasts from each method available on request with approximately 500 copies mailed and with requests still being filled. Also the transformation and precise models for each series when using the Box-Jenkins methodology, (Lusk and Neves, 1984, compared the models used by Andersen, 1983, and found few or no differences), and the Parzen approach (Newton and Parzen., 1984) has been documented and made available to those wishing to evaluate them All of the results have thus been replicated. In addition the results have been scrutinized in a number of follow-up studies. For instance, Simmons (1986) found that the seasonal indices for Naive 2 "were not calculated using the exact procedures that were defined in the M-Competition paper ... and the median absolute percentage error comparative measure was not computed as one might expect it to have been and was not documented as such" (p.457) although he concluded that the results remained unchanged even when his criticisms were taken into account. Under such scrutiny we can conclude with reasonable confidence that the results of the accuracy competitions, in particular the M-Competition, have withstood the requirements for objectivity and replicability.

[5] Lewandowski's *FORSYS* was the only method not fully documented in the description of methods used in the M-Competition, Makridakis *et al.* (1984), as Lewandowski believed *FORSYS* was a proprietary system whose details should be kept secret.

The second important issue is whether empirical accuracy studies are methodologically correct and provide useful information. All the objections raised in the previous section have some substance to them and yet neither singly nor collectively do they seem to undermine the accumulated evidence on performance described in section 2. For instance, the concern about aggregating over different series and time horizons is legitimate. But empirical forecasting accuracy studies are not the only ones that aggregate dissimilar quantities. Corporate accounting adds the value of chairs to that of micro computers and office buildings to calculate a single number called profit. Should this number not be calculated because no wholly satisfactory method is available? Similarly, should we not use automatic methods because of their possible sub-optimality when in actual forecasting situations (for example when making many thousand forecasts a month for production scheduling or inventory control) there is no alternative? If the field of forecasting is to be useful and relevant we have no choice but to aggregate over dissimilar quantities and use automatic methods. Finally, it was argued that the series used in the various empirical studies do not represent a random sample so that we cannot generalize the studies' findings. While accepting the potential importance of this criticism it must also be accepted that the series analyzed cover a wide range of real-life situations and no selection biases have been established. Moreover, the studies include various time spans, countries, companies and levels of aggregation. In addition, new research such as that conducted by Armstrong and Collopy (1993) which uses three large sets of new and different data, and the Makridakis *et al.* (1993) study which also uses new series further confirm the conclusions that we have identified and increase our confidence to generalize them to new situtations. In summary, different researchers, using different methods, and different data sets have reached some broadly compatible conclusions that do not accord with the statistical theory of time series forecasting. The next question to be addressed is how time series statisticians have responded to this evidence.

4. The Impact of Comparative Accuracy Studies

Citation analysis is a well-accepted method of gauging the impact of a published article both in social and hard sciences. Some of the usual objections to its use are less relevant in the field of applied statistics in that research books are relatively rare. Nor are we interested here in only positive citations so that the often expressed concern that an article with which others disagree may still be heavily cited is irrelevant. In Table 3 we show the citations for each of the major studies, over time and categorized by type of journal based on the SCI and SSCI (A detailed listing is available from the authors).

Table 3 leads us towards a number of conclusions. While there was some immediate interest among statisticians in the Newbold & Granger study, this was not sustained or further stimulated by subsequent work. Although in recent years there has been a slight increase in the citations of empirical studies, an examination of the individual citations demonstrates that such an increase has been wholly due to 'forecasters' publishing in the statistical literature. This conclusion directly contrasts with the attitude among management scientists who have responded positively to the M-Competition and the contrast it provides with the Newbold and Granger study.

The *Journal of Time Series Analysis* (a publication specializing in theoretical articles written by academic statisticians) which is also concerned with modelling and forecasting univariate time series is not included in the above table. Only two citations were discovered during the 12 years of its publication to any of the empirical studies examined in this article. Both of these citations were made in simulation studies where it was noted that the simulation evidence conflicted with the empirical evidence of such empirical studies.

For instance, Table 4 shows in more detail the areas that have absorbed the attention of time series statisticians publishing in *JASA*.[6] With ARIMA models as the dominant paradigm the context in which they have been employed has been that of data and error characteristics, multivariate analysis, seasonality and hypothesis testing. Prediction (11% of the published articles) has remained a minority interest despite its importance at the heart of the scientific method. Whatever validity tests were made for the models (and corresponding estimation techniques) have been carried out using simulation techniques only, which of course begs the question of their relevance to real world data. Comparisons with other models (the multiple hypotheses approach) that Chamberlin (1890, 1965) argued was critical to the rapid development of a subject has not been carried out. Out-of-sample comparisons of the accuracy of alternative models were rarely performed throughout the period and were proportionately less in the last decade of research. Yet the focus of most (if not all) time series work is to develop models that are valid over periods beyond the sample on which they are estimated, and perhaps, following Ehrenberg and Bound (1993), to other similar situations. Thus empirical validation, comparative modelling and the choice between alternative models (and methods) seem to have been regarded as unimportant by theoreticians in the field of statistical forecasting.[7]

---

[6] Because of the choice of *JASA* interest in various topics such as non-linear modelling has been understated.
[7] It is fair to add that since 1983 the *Journal of Business and Economic Statistics* has contributed more fully to empirical studies.

	OR & IT	Forecasting	Statistics	Economics	Other Management & Social Sciences	Science & Eng.	Total
Newbold and Granger: 74-78	9	0	15	4	6	1	35
79-83	11	9	6	9	3	5	43
84-91	36	27	9	6	9	5	92
Total	56	36	30	19	18	11	170
Makridakis and Hibon: 79-83	3	0	2	0	1	0	6
84-91	6	6	3	2	0	2	19
Total	9	6	5	2	1	2	25
M-Comp: 82-86	18	48	3	2	8	1	80
87-91	30	59	6	6	14	4	119
84-91	45	87	8	7	22	5	174
Total	93	194	17	15	44	10	373
Meese and Geweke: 84-88	2	2	2	2	0	0	8
84-91	4	2	2	3	2	0	13
Total	6	4	4	5	2	0	21
	Citations Per Year by Subject Area and Total						
74 - 78	1.8	0	3.0	0.8	1.2	0.2	7.0
79 - 83	3.4	6.2	1.8	1.8	1.0	1.0	15.2
84 - 91	11.4	15.2	2.8	2.3	4.1	1.5	37.4
Overall Annual Average	6.5	8.4	2.6	1.8	2.4	1.0	22.8

Table 3 Citations of Accuracy Studies by Subject Area and Year: 1974-1991

N.B. The numbers in the above table are subject to revision.

In summary, the evidence is straightforward: those interested in applying forecasting regard the empirical studies as directly relevant to both their research and to applications in business and non-profit organizations, while those interested in developing statistical models (presumably without concern for their application) pay little attention or ignore such studies.

KEYWORD	%Total Number
Context	
- Data Characteristics, Errors & Distributional Aspects	21%
- Multivariate Analysis	15%
- Seasonality	12%
- Prediction	11%
- Hypothesis Testing	10%
- Diagnostics & Identification	4%
- Parameter Stability	1%
- Others	19%
	100%
Validation	
- Simulation	16%
- Outside Sample	5%
- Robustness	3%
- Diagnostics	2%
	26%
Paper published that did not make any validations	74%
Alternative Models	
- Comparative Forecasting Performance	6%
- Model Comparisons	5%
- Model Identification	3%
	14%
Paper published that did not consider alternative models	**86%**
TOTAL REFERENCES	**241**

Table 4 Context in which Model is Used, Validation Tests Employed and whether Alternative Models are considered: JASA: 1971-1991

N.B. These numbers are subject to revision.

5. Reconciling Theory and Empirical Findings

Kuhn writes "Discovery commences with the awareness of anomaly ... It then continues with a more or less extended exploration of the area of anomaly. And it closes only when the paradigm theory has been adjusted so that the anomalous has become the expected "(p. 52-53). And later he states that once a new paradigm has been embraced "it is rather as if the professional community has been suddenly transported to another planet where familiar objects are seen in a different light and are joined by unfamiliar ones as well" (p.110).

We (and others, Pant and Starbuck 1990, Ord 1986) have argued that the empirical anomalies in time series forecasting identified in section 3 are important and withstand the various criticisms levelled at undermining them. A limited exploration of these issues has also taken place within the management science and forecasting literature (Makridakis, 1986, Pant and Starbuck 1990, Fildes 1992) and by statisticians working within the forecasting literature (Ord 1986, Chatfield 1986)). The need, as we see it, is to adapt the statistical theory of time series modelling by looking at familiar objects in a new light whilst also realizing that unfamiliar concepts must be considered to further advance the field.

The reason for the inability of the established statistical theory to explain the empirical findings is now obvious. In the world of business and economics the assumption of constancy of patterns embedded in time series does not often hold, at least for a good part of the components that constitute the time series. This contrasts with Eherenberg and Bound's (1993) strong claim that law-like relationships are common in the social (business and economic) sciences and such laws are stable across different populations, i.e. both cross-sectionally and across time. Thus, statistical forecasting cannot extrapolate beyond available data assuming that the future will be similar to the past. The evidence we have cited underscores the weakness of assuming (1) the best model/method for post-sample forecasting will be the same as the best-fit model, or (2) the post-sample uncertainty will be compatible with the model based uncertainty. This evidence creates a paradox. If we are certain that the future will differ from the past how can we extrapolate historical information to predict such a future. In our view this paradox can be resolved if we (a) realize that some elements of the future will be similar to those of the past, (b) understand the extent of the differences between the past and the future, and (c) incorporate the possibility of change and its implications in our methods and predictive accuracy/uncertainty. These three aspects serve to focus attention on the structurally stable features of the data, e.g. seasonality, and to detect change and its likely form. By now we know that future

changes in patterns/relationships depend upon the specific data being used, the present state of the economy, industry and firm and the time horizons of the forecast (the longer the horizon, the higher the level of disaggregation, the more likely a change). Moreover we know that some methods are more appropriate than others in dealing with different types of data and possible changes.

The Assumption of Constancy

Sophisticated methods provide, no doubt, a better fit to the data as they are capable of identifying and appropriately estimating complicated patterns. They can provide more accurate forecasts therefore, when the pattern identified is the same as that prevailing during the period being forecasted. In many (if not most) applications, however, three things can happen: First, some recent components in the pattern in the historical data can be temporary (e.g., there is an unexpected/unusual recession and sales fall). In such a case, sophisticated methods will produce inadequate[8] forecasts by extrapolating such temporary patterns as a permanent change and in so doing make large errors. Simpler methods, on the other hand may do better because they are insensitive to such changes. Second, a temporary or permanent change (not part of the historical record) can occur in the future making the forecasts of sophisticated methods inaccurate since they assume no such changes. Here again, simpler methods may do better because they hedge their forecasts by staying close to an "average" pattern. As the chance of some temporary or permanent change increases with the forecasting lead time there is a greater chance of a pattern change and an increased possibility of a larger error if wrong patterns are extrapolated (consider for instance a rising quadratic trend extrapolated twelve periods ahead and consequently, for whatever reasons, such a trend is reversed. Third, special events/actions often unknown to the forecaster such as a competitive promotional campaign may cause fluctuations in the data which appear to have an identifiable and stable structure that can confuse sophisticated methods. These three causes of change may result in less adequate forecasts from sophisticated than simpler methods which are less responsive to such changes. However in the long run temporary changes usually cancel out making methods such as Parzen's ARARMA that use long memory filters more accurate than alternatives that do not.

Improving Time Series Forecasting Methods

As we argued in the introduction the primary purpose of time series forecasting is prediction (and control where possible). Time series methods (with the exception of decomposition models) will rarely be useful for just describing the historical data pattern. Yet, as Table 4 shows the research

[8] The description 'inadequate' is used rather than the more conventional 'inaccurate' to convey the notion that while any or all model based forecasts might be inaccurate, the sophisticated methods were substantially worse than simpler alternatives.

literature has concerned itself with within sample issues such as hypothesis testing (based on sample assumptions that do not hold outside the data being analyzed). Box and Jenkins (1970) and subsequent publications such as Jenkins (1982) stress the importance of working with the client to understand the causes of data fluctuations and bring that understanding to the modeling process. While all time series forecasting must assume that established patterns will not change during the period being forecasted - the 'assumption of constancy' - by focusing sharply on the specific problem in hand and the additional information available to the forecaster it may be possible to overcome some of the constraints this necessary assumption imposes. In the discussion below we comment briefly on various approaches that have the potential for explaining the empirical evidence on accuracy and dealing with the common situation when the the 'assumption of constancy' fails.

Exploiting Robustness: Any chosen model will be robust to a limited range of pattern changes. To improve forecasting accuracy it is necessary for us to know how various methods identify and extrapolate established patterns and how they deal with various types of non-random changes. Linear Regression, for instance, may be better suited to long-term forecasting as it treats all fluctuations around the average pattern (trend) as random. This means that by ignoring the autocorrelation structure in the data, which from a statistical point of view leads to inefficient estimates of the parameters, regression can improve long-term predictions by ignoring an unstable error structure. From the empirical evidence we have cited on comparative performance it would seem that single exponential smoothing or damped trend smoothing have the property of robustness (compared to more general ARIMA alternatives) over a broad range of situations when the trend in the data changes. This is not the case for ARIMA models which concentrate on the autoregressive structure at the expense of the trend, or adaptive methods which attempt to identify and extrapolate the latest error pattern; these are therefore intrinsically better suited to very short-term forecasts when patterns are likely to continue unchanged. Similarly, the covariance of forecast errors from alternative methods does not seem to help in combining their forecasts when it is used to choose the combining weights (Clemen, 1989) because it is unstable over time.

Trend Modelling and Stationarity: Much of the conflicting empirical evidence on comparative forecasting accuracy was concerned with how different models specified trend. Ord (1986) discusses how AR(1) models and IMA(0,1,1) models are poor approximations for each other in longer term forecasting and yet near-equivalent models with no prospect of discriminating between the two (or a linear trend) when a model is selective from in-sample information. Granger (1988) has proposed certain general models that, he argues, should be sufficient to capture stable trends. However, instability in trends could not be included except through the use of multivariate

information which Granger himself admits are not appropriate for the short time typical of the social sciences, at least compared to the physical sciences. Identifying the consequences of mis-specification may help in selection. Alternatively, additional evidence may be necessary to effect the discrimination. Collopy and Armstrong (1993) propose the notion of a qualitative classification of the time series under analysis where short and long term trend are identified and *a priori* evidence on the economic/social determinants of the time series are included to establish compatibility between long and short-term trends. Such an approach is both practical (in that some computer packages used in inventory control already do something similar) and relates directly to the type of information that the analyst would have when forecasting on-line (or analyzing retrospectively). This includes the need to be clear about the problem context in which the time series is to be used.

Lead Time Effects: Empirical research has found relative performance changing with forecast lead time. This has been regarded as an empirical 'fact' for some time with proposals being put forward such as that of Makridakis (1990) that attempt to select different forecasting methods for different lead times. Certainly the error consequences for different methods are affected by lead time as Ord's (1986) example makes clear. Here new theoretical work is being done which, it is hoped, may effectively explain the established empirical results (Findlay 1983, 1985, Bhansali, 1993, Tiao, 1992) and lead to a unified approach to forecasting for different lead times.

Method/Model Selection: In the discussion on robustness we described how notions of likely change (available *a priori* to the analyst) might help in model/ method selection. Similarly, the conservative strategy of 'combining' which has proved so successful has the effect of hedging the forecaster's bets to achieve a second best. Ord (1986) offers a simple illustration of why combining may well work even when there is in fact a best model. However there are substantial gains to be made if the best model *ex post* could be recognised *ex ante* (Fildes, 1989). Research is therefore needed to integrate the model/method choice and combining literature, as combining is another approach to achieving robust forecasts[9].

Multivariate Extensions and Additional Information: In this paper we have concentrated on univariate time series forecasting. However, many areas of application are naturally multivariate from forecasting a number of related products (that may serve the same product market) to more conventional econometric model building. Research on the relative accuracy advantages of multivariate models has been ubiquitous since the early 1970s when Nelson (1972) compared

[9] See Collopy and Armstrong (1992) for an empirically based approach to this issue.

macroeconomic models with ARIMA models. Armstrong (1978) concludes that econometric forecasting does not improve forecasting accuracy while McNees (1990) and Fildes (1985) have concluded that on balance there are advantages to be found from using multivariate methods, although they agree that there are many examples where the converse is true, e.g. Ashley (1988), Huss (1985), Geurts and Kelly (1986), Brodie and De Kluyver (1987), and Arora and Smyth (1990). Much of the evidence on accuracy has been based on *ex post* analysis which as Ashley (1988) points out is suspect if the conditioning variables are themselves poorly forecast as is often the case with macreconomic explanatory variables. In contrast, variables such as temperature in short term forecasting will usually lead to improved performance (Engle *et al.*, 1988), if known in advance which is not unfortunately the case. Theoretical accuracy of multivariate Vector Auto Regressive (VAR) models should be higher than that of the ARIMA or exponential smoothing alternatives. However Riise and Tjostheim (1984) conclude from their investigation that "although in theory multivariate forecasts perform better, in practice one should be careful when trying to implement such models. This is because forecasts based on joint models may be more sensitive than univariate ones to changes in structure." Similar considerations of robustness therefore apply here as they did with univariate methods: research that identifies the circumstances in which multivariate models are robust to mis-specifications in their constituent parts should prove useful. Already, some studies have concluded that the use of loosely informative priors in Bayesian Vector Auto Regressive (BVAR) models are likely to outperform alternatives (Kling and Bessler, 1985; Lupoletti and Webb, 1986). However these studies are concerned with macroeconomic and industy level forecasting situations. At the same time, Makridakis *et al* (1993) examining primarily company based sales forecasting found no improvements in accuracy from using additional information in either a formal or an informal way.

Some progress has been made in building multivariate extensions to ARIMA and exponential smoothing methods, see for example Harvey (1986) and Thisted and Wecker (1981). Here the imposition of a common structure across related time series could well produce the robustness and *ex ante* stability that is wanted. Essentially loose *a priori* knowledge that the series should have common structure is being used, first to identify that structure and second to estimate the parameters perhaps using a Stein type approach. Similarly a common and long-standing feature of many inventory control packages is an assumed common seasonality although only Bunn and Vassilopoulos (1993) have analyzed its performance. However, the extra accuracy that might be obtained through multivariate extensions needs to be weighted against the extra cost and the greater sophistication needed in using these methods. Furthermore, more research and additional empirical studies will be

required to establish the conditions conducive to their predictive superiority, if any, for out-of-sample forecasts in comparison to simple univariate time series alternatives.

6. Conclusions

In this paper we have argued that time series statisticians should pay more attention to the empirical findings on comparative forecasting accuracy and the anomalies that these highlight. Predictive success remains the cornerstone of any science and time series forecasting must strive to achieve that objective to increase its relevance and usefulness. Thus, at first impression it would seem that those theoretical statisticians whose aim is to advance their fields would respond positively to the empirical research. In doing so they could identify new and promising directions for theoretical and applied research and re-orient the time series paradigm away from hypothesis testing and within-sample model fitting to out-of-sample forecasting accuracy. This has not, however, been the case, giving further support to Kuhn's assertion that changing a paradigm is a slow and usually non-evolutionary process. Perhaps a more optimistic note can be struck by the current interest in non-stationary and long memory models, and lead time effects, although interestingly this activity is thought to have arisen independent of the empirical evidence. If so the developments in time series are more in keeping with Lakatos's (1970) view that [a theory] is not falsified ... the existence of many anomalies are not "a sufficient condition for eliminating a specific theory" which cannot be regarded as falsified until a better one emerges. This perspective focuses attention on the lack of engagement between theoreticians and those interested in applications and the need to develop a successful research programme that "explains the previous success of its rival and supersedes it by a full display of heuristic power" where by 'heuristic power' Lakatos means the power of a research programme to anticipate novel facts.

In any joint research programme between those whose primary allegiance is to applications and theoreticians the first issue to be considered are the common benefits that can be obtained by both groups. We believe such benefits are substantial as theoretical statisticians need to focus their research into areas which will potentially (at least in the long run) be useful and relevant to some classes of real-life problem. Similarly, application oriented forecasters need more accurate methods for predicting beyond available data. For example there is a clear need to predict a large number of related time series with parameters updated, say, every period. A theoretical model with measurable performance characteristics and well-defined error bounds that could accomplish such a task will be of great value but is unfortunately not available at present.

We have argued in this paper that empirical studies in the field of forecasting provide a unique opportunity for experimentation that if organized appropriately can also result in replicability and objectivity. Replication can in turn lead to refined guidelines on comparative performance.[10] To this end effective empirical studies can become indispensable tools with considerable potential in guiding the evolution of a research programme that will result in re-defining the dominant paradigm in time series statistics and making it more useful and relevant in the real world.

Although we can show the value of empirical research we are well aware that without a theoretical explanation of observed results on accuracy, the application of those results to new circumstances is akin to driving with no lights and no map. Research such as Collopy and Armstrong's (1993) on rule-based forecasting and research on combining where no explicit stochastic model is proposed can be criticized on this count.

The aim of our paper has been to increase the awareness of the empirical findings on accuracy with the intention of encouraging theoretical statisticians to develop an appropriate theory and improved methods to respond to the anomalous evidence. For application-oriented and empirically-based researchers (such as ourselves) the need to develop a theoretical framework, improve methods and selection criteria when the objective is post-sample forecasting accuracy is a necessity. Moreover, we believe that a research program would be best carried out through a joint effort between those interested in applications and theory. Otherwise, we believe that progress will be slow and "costly" to both groups.

[10] This notion is much more modest than that proposed by Ehrenberg where the laws he claims are common in the social sciences will typically be replicable and "any serious exception [ ]would rate as a scientific discovery (a mini-paradigm shift) and not merely a failure to predict". He states that such an event "does not happen often or easily".

REFERENCES

Andersen, A. (1983) An empirical examination of Box-Jenkins forecasting , *Journal of the Royal Statistical Society (A)*, **145**, 472-475.

Anderson, O.D. (1979) Comment on Makridakis, S. and Hibon, M., Accuracy of forecasting: an empirical investigation with discussion, *Journal of the Royal Statistical Society (A)*, **142**, 134.

Armstrong, J. S. (1978) Forecasting with econometric methods: folklore versus fact with discussion, *Journal of Business*, **51**, 549-600.

Armstrong, J.S. (1979) Advocacy and objectivity in science, *Management Science*, **25**, 423-428.

Armstrong J. S. and Lusk E. J. (eds.) (1983) Commentary on the Makridakis time series competition (M-Competition), *Journal of Forecasting*, **2**, 259-311.

Armstrong, J. S. and Collopy, F. (1992) Error measures for generalizing about forecasting methods: empirical comparisons with discussion, *International Journal of Forecasting*, **8**, 69-80, 99-.

Armstrong, J. S. and Collopy, F. (1993) Causal forces: Structuring knowledge for time series extrapolations, *Journal of Forecasting*, **12**, 103-115.

Arora, H.K. and Smyth, D.J. (1990) Forecasting the developing world: an accuracy analysis of the IMF forecasts, *International Journal of Forecasting*, **6**, 393-400.

Ashley. R. (1988) On the relative worth of recent macroeconomic forecasts, International *Journal of Forecasting*, **4**, 363-376.

Box, G.E.P. and Jenkins, G.M. (1973) Some comments on a paper by Chatfield and Prothero and on a review by Kendall (with a reply), *Journal of the Royal Statistical Society (A)*, **136**, 337-352.

Box, G.E.P. and Jenkins, G. (1976) *Time Series Analysis, Forecasting and Control* (2nd ed.), Holden-Day, San Francisco.

Bhansali, R.J. (1993) Fitting autoregressive models for multistep prediction, University of Liverpool

Bramson, M.J. (1974) Comment on Newbold, P. and Granger, C.W.J., Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society (A)*, **137**, 157.

Brodie, R. J. and De Kluyver, C. A. (1987) A comparison of the short-term forecasting accuracy of econometric aand naive extrapolation models of market share with discussion, *International Journal of Forecasting*, **3**, 423-462.

Brown, R.G. (1962) *Smoothing, forecasting and prediction*, Prentice-Hall, Englewood Cliffs, N.J..

Bunn, D. W. and Vassilopoulos, A.I. (1993) Using group seasonal indices in multi-item short-term forecasting, *International Journal of Forecasting*, **9**, forthcoming.

Chamberlin, T.C. (1965) The method of multiple working hypotheses (originally published in Science in 1890), *Science*, **148**, 754-759.

Chatfield, C. and Prothero, D.L. (1973) Box-Jenkins seasonal forecasting: problems in a case study with discussion, *Journal of the Royal Statistical Society (A)*, **136**, 295-336.

Chatfield, C. (1978) The Holt-Winters forecasting procedure, *Applied Statistics*, **27**, 264-279.

Chatfield, C. (1988) What is the 'best' method of forecasting? *Journal of Applied Statistics*, **15**, 19-38.

Chatfield, C. (1993) Calculating interval forecasts, *Journal of Business and Economic Statistics*, in press.

Clemen, R. (1989) Combining forecasts: A review and annotated bibliography with discussion, *International Journal of Forecasting*, **5**, 559-608.

Close, F. (1990) *Too Hot to Handle: The Story of The Race for Cold Fusion*, W.H. Allen, London

Cogger, K.O. (1974) The optimality of general-order exponential smoothing, *Operations Research*, **22**, 858-867.

Collopy, F. and Armstrong, J. S. (1992) Rule-based forecasting, *Management Science*, **38**, 1394-1414.

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

De Gooijer, J.G., Abraham, B, Gould, A and Robinson L. (1985) Methods for determining the order of an autoregressive-moving average process: A survey, *International Statistical Review*, **53**, 301-329.

De Gooijer, J. G., and Kumar, K. Some recent developments in time series moodelling, testing and forecasting, *International Journal of Forecasting*, **8**, 135-156.

DeJong, D. N. et. al. (1991) Integration versus trend stationarity in time series, Econometrica, 423-434.

Dewald, W.G., Anderson, R.G. and Thursby, J.G. (1986) Replication in empirical economics: the Journal of Money, Credit and Banking project, *American Economic Review*, **76**, 587-603.

Durbin, J. (1979) Comment on: Makridakis, S. and Hibon, M."Accuracy of Forecasting: An empirical investigation", *Journal of the Royal Statistical Society (A)*, **142**, 133-134.

Ehrenberg, A.S.C. and Bound, J.A. (1993) Predictability and prediction, *Journal of the Royal Statistical Society (A)*, in press .

Engle, R. F., Brown, S. J. and Stern, G. (1988) A comparison of adaptive structural forecasting methods for electricity sales, *Journal of Forecasting*, **7**, 149-172.

Ferber, R. (1956) Are correlations any guide to predictive value, *Applied Statistics*, **5**, 113-122.

Fildes, R. (1985) Quantitative forecasting - the state of the art: econometric models, *Journal of the Operational Research Society*, **36**, 549-580.

Fildes, R. (1983) An evaluation of Bayesian forecasting, *Journal of Forecasting*, **2**, 137-150.

Fildes, R. and Makridakis, S. (1988) Loss functions and forecasting, *International Journal of Forecasting*, **4**, 545-550.

Fildes, R. (1989) Evaluation of aggregate and individual forecast method selection rules. *Management Science*, **35**, 1056-1065.

Fildes, R. and Beard, C. (1992) Forecasting systems for production and inventory control, *International Journal of Operations and Production Management*, **12**, 4-27.

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

Findley, D.F. (1985) Model selection for multi-step-ahead forecasting. Identification and System Parameter Estimation , In *7th IFAC/IFORS Symposium*, 1039-1044.

Gardner, E. S. Jr., and McKenzie, E. (1985) Forecasting trends in time series, *Management Science*, **31**, 1237-1246.

Geurts, M.D. and Kelly, J.P. (1986) Forecasting retail sales using alternative models, *International Journal of Forecasting*, **2**, 261-272.

Geweke, J. and Porter-Hudak, S. (1983) The estimation and application of long memory time series models, *Journal of Time Series Analysis*, **4**, 221-238.

Granger, C.W.J. (1988) Models that generate trends, Journal of Time Series Analysis, **8**, 329-343.

Grenander, U. and Rosenblatt, M. (1957) *Statistical analysis of stationary time series*, Wiley, New York.

Hannan, E.J. (1960) *Time Series Analysis*, Methuen, London.

Harrison, P.J. (1967) Exponential smoothing and short-term sales forecasting, *Management Science*, **13**, 821-842.

Harvey, A. C. (1986) Analysis and generalization of multivariate exponential smoothing, *Management Science*, **32**, 374-380.

Hill, G. and Fildes, R. (1984) The accuracy of extraplation methods: An automatic Box-Jenkins Package SIFT, *Journal of Forecasting*, **3**, 319-323.

Huss, W.R. (1985) Comparative analysis of company forecasts and advanced time series tecyhniques in the lectric utility industry, *International Journal of Forecasting*, **1**, 217-239.

Jenkins, G.M. (1974) Comments on Newbold, P. and Granger, C.W.J. "Experience with forecasting univariate time series and the combination of forecasts", *Journal of the Royal Statistical Association*, **137**, 148-150.

Jenkins, G.M. (1982) Some practical aspects of forecasting in organizations, *Journal of Forecasting*, **1**, 3-21.

Johnstone, D.J. (1989) On the necessity of random sampling, *British Journal of the Philosophy of Science*, **40**, 443-483.

Kling, J. L. and Bessler, D. A. (1985) A comparison of multivariate forecasting procedures for economic time series, *International Journal of Forecasting*, **1**, 5-24.

Koehler, A. B. and Murphree, E. S. (1988) A comparison of results from state space forecasting with forecasts from the Makridakis Competition, *International Journal of Forecasting*, **4**, 45-55.

Kramer, W. H. et al. (1985) Diagnostic checking in practice, *Review of Economics and Statistics*, **67**, 118-123.

Kuhn, T.S. (1962) *The structure of scientific revolutions*, University of Chicago Press, Chicago.

Lakatos, I., Falsification and the methodology of scientific research programmes in *Criticism and the Growth of Knowledge, Lakatos I* and Musgrave, A. (eds.)

Libert,G. (1983) The M-Competition with a fully automatic Box-Jenkins procedure, *Journal of Forecasting*, **2**, 325-328.

Lupoletti, W.M. and Webb, R. H. (1986) Defining and improvingthe accuracy of macroeconomic forecasts: contributions from a VAR model, *Journal ofBusiness*, **59**, 263-.

Lusk, E.J. and Neves, J.S. (1984) A comparative ARIMA analysis of the 111 series of the Makridakis competition, *Journal of Forecasting*, **3**, 329-332.

Makridakis, S. (1976) A Survey of Time Series, *International Statistical Review*, **44**, 29-70

Makridakis, S. and Hibon, M. (1979) Accuracy of forecasting: an empirical investigation with discussion, *Journal of the Royal Statistical Society (A)*, **142**, 97-145.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, P., and Winkler, R. (1982) The accuracy of extrapolation (time series) methods; Results of a forecasting competition, *Journal of Forecasting*, **1**, 111-153.

Makridakis, S. (1983) Empirical evidence versus personal experience in Armstrong, J.S. and Lusk, E.J. Commentary on the Makridakis time series competition (M-Competition), *Journal of Forecasting*, **2**, 295-309.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, P., and Winkler, R. (1984) *The Forecasting Accuracy of Major Time Series Methods*. Wiley, Chichester, UK.

Makridakis, S. (1986) The art and science of forecasting; an assessment and future directions, *International Journal of Forecasting*, **2**, 15-39.

Makridakis, S., Hibon, M. Lusk, E. and Belhadjali, M. (1987) Confidence intervals: an empirical investiagation of the series in the M-competition, *International Journal of Forecasting*, **3**, 489-508.

Makridakis, S. and Winkler, R. L. (1989) Sampling distributions of post-sample forecasting errors, *Applied Statistics*, **38**, 331-342.

Makridakis, S. (1990) Sliding Simulation: A New Approach to Time Series Forecasting, *Management Science*, **36**, 505-512

Makridakis, S. and Hibon, M. (1991) Exponential smoothing: the effect of initial values and loss functions on post-sample forecasting accuracy, *International Journal of Forecasting*, **7**, 317-330.

Makridakis, S. et al. (1993) The M-2 Competition: A real-life judgmentally based forecasting study, *International Journal of Forecasting*, forthcoming.

McNees, S. K. (1990) The role of judgment in macroeconomic forecasting accuracy.*International Journal of Forecasting*, **6** 287-299.

Meese, R. and Geweke, J. (1984) A comparison of autoregressive univariate forecasting procedures for macroeconomic time series, *Journal of Business and Economic Statistics*, **2**, 191-200.

Nelson, C.R. (1972) The prediction performance of the FRB-MIT-PENN model of the US economy, *American Economic Review*, **62**, 902-917.

Nelson, C.R. and Plosser, C.I. (1981) Trends and random walks in macroeconomic time series: Some evidence and implications, *Journal of Monetary Economics*, **10**, 139-162.

Newbold, P. and Granger, C.W.J. (1974) Experience with forecasting univariate time-series and the combination of forecasts with discussion, *Royal Statistical Society (A)*, **137**, 131-165.

Newbold, P. (1983) The competition to end all competitions in Armstong, J. Scott and Lusk, Edward J., (eds), Commentary on the Makridakis time series competion (M-Competition), *Journal of Forecasting*, **2**, 276-279.

Newton, H. J. and Parzen, E. (1984) In *Forecasting and time series model types of 111 economic time series in Makridakis, Spyros et al, The Forecasting Accuracy of Major Time Series Methods*, Wiley, Chichester, England.

Ord, J. K. (1988) Future developments in forecasting: The time series connexion, International *Journal of Forecasting*, **4**, 389-401.

Pant, P. N. and Starbuck, W. H. (1990) Innocents in the forecast: forecasting and research methods, *Journal of Management*, **16**, 433-460.

Parzen, E. (1982) ARARMA models for time series analysis and forecasting, *Journal of Forecasting*, **1**, 67-82.

Pierce, D.A. (1977) Relationships - and the lack thereof -between economic time series, with special reference to money and interest rates, *Journal of the American Statistical Association*. **72**, 11-26.

Popper, R.K., 1972, *Objective Knowledge*, Oxford University Press, Oxford.

Poulos, L., Kvanli, A. and Pavur, R. (1987) A comparison of the accuracy of the Box-Jenkins method with that of automated forecasting methods, *International Journal of Forecasting*, **3**, 261-267.

Priestley, M.B. (1974) Comment on Newbold, P. and Granger, C.W.J., "Experience with forecasting univariate time series and the combination of forecasts", *Journal of the Royal Statistical Society (A)*, **137**, 152-153.

Priestley, M.B. (1979) Comment on Makridakis, Spyros, Hibon, Michelle, Accuracy of forecasting: an empirical investigation with discussion, *Journal of the Royal Statistical Society (A)*, **142**, 127-128.

Prothero, D.L. and Wallis, K.F. (1976) Modelling macroeconomic time series with discussion, *Journal of the Royal Statistical Society (A)*, **139**, 468-500.

Reid, D.J (1969) A comparative study of time series prediction techniques on economic data, University of Nottingham, Nottingham.

Reid, D. J. (1972) A comparison of forecasting techniques on economic time series in Forecasting in Action, Bramson, M.J., Helps, I.G. and Watson-Gandy, J.A.C.C. (eds.), *Operational Research Society*, Birmingham, UK.

Reid, D.J. (1974) Comment on: Newbold, P., Granger, C.W.J., Experience with forecasting univariate time-series and the combination of forecasts with discussion, *Journal of the Royal Statistical Society (A)*, **137**, 146-148.

Riise, T. and Tjostheim, D. (1984) Theory and practice of multivariate ARIMA forecasting, *Journal of Forecasting*, **3**, 309-317.

Rosenberg, A. (1992) Economics - Mathematical Politics or Science of Diminishing Returns, Chicago U.P., Chicago.

Schnaars, S. P. (1986) A comparison of extrapolation models on yearly sales forecasts, *International Journal of Forecasting*, **2**, 71-85.

Schupack, M.P. (1962) The predictive accuracy of empirical demand analysis, *Economic Journal*, **72**, 550-575.

Simmons, L.F. (1986) M-Competition - a closer look at NAIVE2 and Median APE: a note, *International Journal of Forecasting*, **4**, 457-460.

Texter, P. A., Ord, J. K. (1989) Forecasting using automatic identification procedures: A comparative analysis, *International Journal of Forecasting*, **5**, 209-215.

Thisted, R.A. and Wecker, W.E. (1981) Predicting a multitude of time series, *Journal of the American Statistical Association*, **76**, 516-523.

Tiao, G. C. and Xu, D. (1991) Robustness of MLE for mutli-step predictions: the exponential smoothing case, Technical Report No. 117, University of Chicago, Chicago.

Watson, M. W. (1986) Univariate detrending methods with stochastic trends, *Journal of Monetary Economics*, **18**, 49-75.

Watson, M.W., Pastuszek, L. M. and Cody, E. (1987) Forecasting commercial electricity sales, *Journal of Forecasting*, **6**, 117-136.

Williams, W.H. and Goodman, M.L. (1971) A simple method for the construction of empirical confidence limits for economic forecasts, *Journal of the American Statistical Association*, **66**, 752-754.

Wold, H. (1938) A Study in the Analysis of Stationary Time Series, Almgrist & Wiksell, Stockholm

Wu, L.S.-Y., Ravishaker, N., and Hosking, J.R.M. (1991) Forecasting for business planning: a case study of IBM product sales, *Journal of Forecasting*, **10**, 579-595.

Zellner, A. (1986) A tale of forecasting 1001 series: The Bayesian knight strikes again, *International Journal of Forecasting*, **2**, 491-494.