Bibliography on business and economic forecasting, 2014

Contents

Introduction ................................................................................................................................................................... 3
Textbooks and Overviews .............................................................................................................................................. 4
Journals .......................................................................................................................................................................... 5
Forecasting Software and Support Systems .................................................................................................................. 6
General Articles and Commentaries ............................................................................................................................ 7
Judgmental Methods ..................................................................................................................................................... 8
  Formalizing Judgment .......................................................................................................................................... 10
  Bootstrapping Judgments .................................................................................................................................... 11
  Improving Judgment ............................................................................................................................................ 12
Extrapolative Forecasting ............................................................................................................................................ 13
  Exponential Smoothing ........................................................................................................................................ 14
  Autoregressive Integrated Moving Average Models ........................................................................................... 15
  State-space Models ............................................................................................................................................. 16
Econometric Forecasting ............................................................................................................................................. 17
  Econometric Forecasting Books ........................................................................................................................... 17
  Commentary on the State of the Art ....................................................................................................................... 18
  Unit Root Testing ............................................................................................................................................... 19
  Vector Autoregressions ....................................................................................................................................... 21
  Cointegration ...................................................................................................................................................... 21
Non-linear Computer-intensive Methods .................................................................................................................. 23
Comparing and Combining Forecasts .......................................................................................................................... 25
  Forecasting Competitions .................................................................................................................................. 26
  Measurement of Forecast Accuracy ....................................................................................................................... 28
  Comparison of Forecast Errors ........................................................................................................................... 29
  On Selecting the Best Model ............................................................................................................................... 30
  Combining ............................................................................................................................................................ 32
  Encompassing ...................................................................................................................................................... 34
  Forecasting Distributions ..................................................................................................................................... 35
Forecasters and Forecasting Practice ........................................................................................................................... 36
Applications ................................................................................................................................................................. 38
Introduction

Forecasting covers a wide range of activities, even when the focus, as here, is on management, business and economics. Its history goes back at least as far as the Oracle at Delphi in Greece. Stripped of its mystique this was what we today call unaided judgment, the only forecasting method available for centuries. However, as a formal area of study, the earliest examples are from the weather forecasters of the 19th C. The early years of the 20th C saw increasing interest in business and economic forecasting with novel methods applied to agricultural yields and prices as well as analysis of the business cycle.

Alongside an interest in methods of forecasting, researchers became interested in the producers of forecasts who might or might not use formal methods. Construction of macroeconomic models occurred in the 1940 associated with the work of the Cowles Commission. As the first computers became available these models were estimated and used for forecasting. The 1970s marked the era of univariate forecasting; ARIMA modeling and exponential smoothing both date from this time and are widely used today in business forecasting. Up to this time, there was often little exchange between the different groups of researchers, economists, statisticians, engineers and, later, psychologists. The 1980s saw an increasing institutionalization of the subject with the founding of first the Journal of Forecasting and subsequently the International Journal of Forecasting, both of which aimed to integrate the disparate aspects of forecasting. In part this was a recognition that the practice of forecasting, ubiquitous across government, commerce and industry did not recognize the artificial barriers between academic disciplines. It was also the era of some major forecasting developments: unit-root testing, vector autoregression, cointegration, state-space modeling, ARCH modeling all became popular around this time. Questions about the best forecasting method were tested in competitions between methods, but clear answers were not forthcoming.

As computing power has continued to increase, more sophisticated forecasting methods have emerged, in particular the method of neural networks. As to the future of forecasting, it seems a safe bet that we will see computing power used to handle more data, more non-linear methods, more emphasis on forecast distributions and perhaps more on the limitations of business and economic forecasting and the strategies that should be followed. Allen and Fildes give a fuller history of the subject’s development which is drawn on here (Allen and Fildes 2011).

After introducing books and papers that examine the breadth of forecasting this bibliography’s structure recognizes that the fundamental methods of business and economic forecasting: judgment, extrapolative time series methods and econometrics still have distinct development paths. In addition, the newer area of computational intensive methods, often applied to predicting individual consumer behavior is included separately. Forecasting software is also surveyed here briefly as its dissemination has been critical both to research innovation and to changes in forecasting practice.

Finally the bibliography includes various important areas of application, which have also stimulated methodological developments
Textbooks and Overviews

Books on forecasting range widely from the specialist and historical to popular overviews of the subject such as, The Signal and the Noise (Silver 2012). Silver’s expertise in US election forecasting generated massive publicity as to the importance of identifying the useful information in data (the signal) and eliminating randomness (noise). Effectively the book proposes an argument that valuable forecasts are best produced through the careful analysis and modelling of data. Undergraduate or masters text books come in two flavors, generalist and those that focus primarily on a particular set of methods such as time series. Generalists include Makridakis, Wheelwright and Hyndman 1998 and Wilson, Keating, with John Galt Solutions 200) and more recently Hyndman and Athanasopoulos (online) and Ord and Fildes 2013. Diebold 2007 offers a more theoretical perspective. Armstrong’s 2001 book is very different, a collection of chapters written by leading forecasters, built around a set of principles and intended as a reference and guide to researchers and practitioners.

Consists of thirty chapters by some 40 different authors covering all aspects of qualitative and quantitative forecasting, reviewing methods, evaluation, applications, mostly written in a non-technical style. Built around a set of principles, usually supported with evidence of their effectiveness. The principles, and any new additions, are listed on the companion website http://forprin.com. A core reference book.

Advanced undergraduate level, requiring some background in statistics. Covers graphical analysis, various univariate methods, cycles, some limited discussion of regression and the more advanced topics of unit roots and volatility measurement. Includes instructor resources.

Hyndman, Rob J., and George Athanasopoulos. Forecasting Principles and Practice. (http://otexts.com/fpp/)
Available online only. For undergraduates and practitioners with knowledge of basic algebra. Many real-world examples. Uses R for analysis and emphasizes graphical presentation. Covers judgmental forecasting, regression, time-series decomposition, smoothing, ARIMA models and advanced forecasting methods.

This book was one of the first to cover forecasting as it is practiced. It does not require much statistical background beyond an introductory course.
For undergraduates, specialist masters and practitioners. Requires limited statistics background. General principles guide and simplify forecasting practice. Uses real data sets and a variety of software. Covers extrapolative methods (exponential smoothing, ARIMA), multiple regression, data mining and judgmental methods. Discusses forecasting applications in operations and marketing, organizing forecasting support system and dealing with uncertainty. Includes instructor resources.

An entertaining introduction to various important topics in forecasting, many of which Silver has worked on, including elections, poker, weather, climate and economics. Contains many sensible principles but no modelling; in essence it is an argument for careful data analysis and modelling.

Introductory text book that includes a standard forecasting methods but also an introduction to data mining. Number of mini-cases. Has limited discussion of regression modelling and no discussion of judgmental methods. Relies on Galt’s software solution, ForecastXTM, a student edition of which is included.

Journals

The two specialized forecasting journals are the International Journal of Forecasting and the Journal of Forecasting. Forecasting articles also appear in more general methodological journals, in particular the Expert Systems with Applications, Journal of Business and Economic Statistics, and Technological Forecasting and Social Change. In addition particular application areas such as finance and transport discuss various special features of the areas, both substantive and methodological. Forecasting is unusual in that there is much organizational involvement in its practical aspects. Two journals have been established to support and educate forecasting practitioners, Foresight and the Journal of Business Forecasting.

Focused on computationally intensive methods such as neural networks. Weak on the evaluation of the new methods it discusses. Publishes 18 issues per year and first appeared in 1990.

Foresight: The International Journal of Applied Forecasting  http://forecasters.org/foresight/
Forecasting Software and Support Systems

Most formal model-based forecasting is carried out using software, whether it be in-house Excel-based or part of a large-scale Enterprise Resource Planning (ERP) software system. Most forecasting appears to be carried out using Excel, often with unsatisfactory results (Sanders and Manrodt 2003). Kusters, McCullough, and Bell 2006 provide a review of the different types of forecasting systems. Because of the large number of products and their range it is impossible to provide a helpful review. Ord and Fildes (2013, Appendix C, see *Textbooks and Overviews*) examine features in ten packages and also discuss the key dimensions to be considered when purchasing (2013, Chapter 12, p. 398). From time to time reviews of particular packages appear with Yurkiewicz 2013 publishing a biennial review, the latest considering 25 packages. Of note is the freeware package supported by the R community (Hyndman and Khandakar 2008)Software is seldom designed to match organizational forecasting practices. Frequently, design features fail to recognize that forecasters require support to effectively choose between data models, and to add in market information excluded from the available statistical models (Fildes, Goodwin, and Lawrence 2006). There is current research aimed at improving software design (Lee et al. 2007) based on improving the effectiveness of judgmental interventions.


Defines forecasting support systems (FSSs) as model based forecasting and data management systems integrated with the forecast users. First examines what features are typically included (and missing)
from such systems and using research on judgment, then recommends new design features that should make forecasting more organizationally effective.

Describes the R forecasting package and the algorithms employed.

Gives a historical perspective on forecasting software. Defining two different types of application, research oriented and business software to support demand planning they then look at likely future developments.

How can potentially useful information be made effective? Describes an experiment to improve the incorporation of promotional information into a baseline statistical forecast. Offering the forecaster analogous information summarising past promotions with similar features to the forthcoming promotion proved most effective.

Based on a survey of 240 US corporations, shows the majority (60%) of users are dissatisfied with their forecasting software, particularly if it is spreadsheet based. And those who use commercial software rather than in-house seem to achieve lower errors!

Yurkiewicz, J. 2013. Software Survey: Forecasting - An Upward Trend? *OR/MS Today*, no. 39 (3 (June)).
The latest forecasting software survey – but it suffers necessarily from superficiality as testing software is a time consuming business with many dimensions from calculation integrity to report writing facilities. Many established products, particularly in the demand planning area, are not included.

**General Articles and Commentaries**

This section contains reflections by some of the key founding researchers in the discipline of forecasting as to the success achieved, current limitations and the potential for further improvement. De Gooijer and Hyndman 2006 looked back over the 25 years during which forecasting became increasingly established as a research area, examining the achievements in time series forecasting and identifying research gaps. In fact the ambitious aspirations of early researchers in the field to develop new methods which would consistently identify the signal with reliable estimates of the residual noise so that organizational planning could take this into account, were soon confounded. After surveying the evidence on the relative performance of different methods, Makridakis 1986 saw the key problem as the lack of constancy – people and economic structures change unexpectedly and these changes cannot
usually be included in any model. Armstrong 2006 updated the evidence on relative performance and remains skeptical as to the likely success of more complex techniques, but he does identify research initiatives, both in new methods and new approaches he believes likely to pay off. Much of forecasting research and applications, not least exponential smoothing, has fallen into the domain of operations research; Fildes, et al. 2008 surveyed the field from the point of view of applications, arguing that the latest technical developments are not particularly likely to lead to substantial improvements in accuracy, except when they are closely linked to particular applications (see, for example, *Applications*).


Armstrong offers a very personal and stimulating summary of which methods have proved successful in forecasting, providing evidence to support his views.


This non-technical paper, by editors of the journal, is an ambitious attempt to survey 25 years of time-series forecasting (also touching on multivariate econometric model building). Whilst necessarily superficial it offers a panoramic view of the topic and concludes with comments on future research directions.


Evaluates the last 25 years of forecasting research, noting the rapid development of computer intensive methods, in the context of applications in operations and marketing. Concludes by arguing that future developments are likely to be problem specific and that the days of global solutions are finished.


One of forecasting’s most innovative researchers surveys the research evidence on comparative forecasting accuracy, the accomplishments in the field and its shortcomings. Shows how forecasting models can overfit the data; simpler methods are more robust. Based on empirical evidence, the factors affecting accuracy and the barriers to successful implementation. Identifies future research needs, many still relevant today.

**Judgmental Methods**

Non-quantitative judgmental methods of forecasting include individuals or groups agreeing a forecast, role-playing, use of analogies and survey-based approaches. Judgmental forecasting is a relatively neglected academic area, yet it is probably the approach most widely used by practitioners. Judgment is used in almost any forecasting situation, sometimes combined with a formal model. Individual or group judgmental forecasts are perhaps the most common type of forecasts. However, based as they are on simple heuristics, they suffer from various biases (Kahneman and Lovallo 1993; Tversky and Kahneman
Various approaches have been developed to overcome some of these biases, which typically lead to inaccurate forecasts. The Delphi method (developed by the Rand Corporation in 1944) is perhaps the best known. Role-playing, most familiarly seen in war games, can also be used in competitive business situations to structure the interactions between players. Players are instructed in the roles they are representing, are given a description of the scenario and are asked to make decisions. The decisions function as forecasts. Analogies are often used to forecast the fortunes of a start-up new product or service, by finding a similar products that already has a recorded history. If the products are similar, the history of the older product is the basis of a forecast for the start-up. Psychologically-based approaches rely on either unaided or aided expert judgment. At one extreme, the output of large-scale macroeconomic models is adjusted before release to take into account information not used in the model or not available in a form that could have been used. At the other extreme, sales forecasts may be produced with only a simple extrapolative forecast of the product’s history which is then adjusted to take into account the many pieces of information, such as promotions, excluded from the statistical model. Researching the effectiveness of such judgmental adjustments is important both in practice but also to theoretical understanding of how forecasters process information and the heuristics they use. Limitations on human judgment are well-known but their importance is perhaps not widely appreciated. Humans have limited processing power, are subject to biases, and are inefficient decision makers and forecasters. In practical applications such biases can be very damaging to forecast accuracy. Research has attempted to measure and overcome these problems, recommending procedures to improve accuracy. Lawrence et al. 2006 have provided a useful review of the many issues in understanding and improving judgmental forecasting. Three branches of research into judgment can be distinguished: (1) that concerned with methods of formalizing judgment to produce more accurate forecasts than the ‘raw’ judgments alone, (2) modelling experts’ ability to forecast (called Bootstrapping), and (3) measuring and compensating for judgmental biases.

Uses the fundamental ideas from Tversky and Kahneman 1974 to consider whey bold forecasts are often made which seem to discount likely risks but result in poor outcomes. Also considers how judgmental forecasts might be improved by taking an ‘outsider’s view’.

Comprehensive review. Over 200 references. Topics include: unaided judgment, influence of domain knowledge, probability forecasts and prediction intervals, improving via feedback, decomposition, combining, bootstrapping, and suggestions for future research.

Perhaps the key article in the area of judgmental forecasting though it has no direct reference to the topic. Cited over 19000 times, it identifies three fundamental heuristics adopted when judgments are made under uncertainty: ‘representativeness’, ‘availability’, and ‘anchoring’ and adjustment’.
Formalizing Judgment

With judgmental forecasts subject to continual biases and inefficiencies, an important question to ask is how they can be improved. An example of formalizing judgment is the Delphi technique developed by the Rand Corporation in the 1940s aimed at overcoming the limitations of individual or committee judgments. It consists of four components: anonymity, iteration, controlled feedback, and statistical aggregation of group response. Studies using the Delphi technique are generally more accurate than unstructured approaches (Rowe and Wright 1999). Recommendations on how to get the best from a Delphi study are given in Rowe and Wright 2001. Consumer and business surveys where potential customers are surveyed have long been used in forecasting and methods have been proposed to improve the forecasting accuracy of such purchase intentions. However, taking the results of a survey on face value would be unwise – respondents don’t tell the truth as they know it, nor do they know their future circumstances, which will influence their purchasing decisions. Morwitz, et al. 2007 propose using consumer segmentation to get the most information from intentions surveys. Forecasting the outcome of conflicts, whether political or in the economy, is important and dominated by the opinion of experts. Tetlock 2005 examined more than 80,000 political and economic forecasts by experts, found them wanting and proposed explanations for their inadequacy, for example, the collapse of communism in 1989. Green 2002 evaluated two more formal approaches to forecasting conflicts, the Delphi and Role playing, comparing them to unaided judgment. Role playing, commonly used in war gaming, proved the most effective with unaided judgment the least. But the claims made by game theorists for the practical utility of their models was also cast into doubt. A commentary on the article presents alternative views of Green’s experimental design.


Notes that variations in application of the technique make generalizations about Delphi difficult. Reviewed empirical studies suggest that Delphi groups outperform statistical groups and standard interacting groups, although there is no consistent evidence that the technique outperforms other structured group procedures.

Tetlock, Philip E. 2005. Expert political judgment: How good is it? How can we know? Princeton, NJ: Princeton University Press. What do experts really have to offer? Are some experts better than others? Examines 80,000+ political and economic forecasts by 284 experts. Experts suffer from all the problems inherent in judgment and their performance proved poor, and the greater their expertise the worse their record. Experts with a cautious analytic approach do better.

**Bootstrapping Judgments**

Because judgments, made by both individuals or groups, tend to be biased and inefficient, one approach where there is a set of past judgmental forecasts available is to develop a statistical model of these forecasts – a so-called bootstrap model. This notion has been applied in a wide range of cross-sectional applications from examining clinical judgments to financial and legal forecasts. The result is invariably a more accurate forecast based on the model – a model of the judge – than the raw forecasts (or judgments) themselves. Dawes, et al. 1989 surveyed a number of applications from the medical field. Karelaia and Hogarth 2008 examined a large number of studies and offer insight into where bootstrap models are likely to outperform raw judgment. The evidence used in these studies is cross-sectional but in forecasting applications which include an explicit time dimension we need to be convinced that such bootstraps are also valuable. Fildes 1991 showed how this can be done to improve on standard methods of combining statistical models with bootstraps. He also pointed out that for time series, bootstraps are not inevitably successful.


Unaided judgment preferred to statistical recommendations – but why? Examines primarily medical evidence concluding that statistical diagnostic models outperform the judgment of clinicians. The authors offer suggestions on why this discrepancy occurs. Unaided judgments are used extensively in business – could and should statistical models be embraced more?


A panel of construction industry experts produced regular forecasts of future construction output based on many factors but including an explicit forecast of GDP. Study shows that a raw bootstrap of the judgmental forecasts is inadequate and proposes an error bootstrap which eliminated the biases in the bootstrap model and produces more accurate forecasts.


Provides a wide-ranging review of bootstrap models and the conditions in which they are expected to outperform raw judgments.
Improving Judgment

Judgmental forecasting sees the forecaster as interpreting various sources of information and advice which are then combined (usually without resort to any explicit weighting scheme) to deliver the required forecast. This process is subject to what has come to be called ‘heuristics and biases’. However, judgment is often used in combination with model-based forecasts. Bunn and Wright 1991 examined the link between judgment and formal statistical methods of forecasting showing the value of judgmental approaches despite the limitations of bias and inconsistencies identified by Tversky and Kahneman, 1974, and Kahneman and Lovallo 1993. (See *Judgmental Methods*.) They presented evidence to contradict what was for a while the accepted wisdom that judgment is inevitably inferior to feasible statistical methods. Further evidence from the need to estimate software development effort was given by Jorgensen 2007. Goldstein and Gigenrenzer 1990 provided evidence that some simple heuristics work well. For example, Blattberg and Hoch 1990 proposed a simple method of averaging, though as Fildes, et al. 2009 have recently pointed out this is not a catch-all solution and propose better models. Lawrence, et al. 1986 provided experimental evidence on a variety of methods of combination. One of their conclusions is that the combination should be calculated through an explicit formula for the combination rather than relying on judgments for the combination. This article is a good example of how experiments can be used effectively to draw conclusions relevant to forecasting in practice. Often forecasts are produced based on other people’s recommendations. The question this raises is how much the recipient of the information and advice trusts and relies on the source and in what circumstances. Harvey and Fischer 1997, in another experimental study, modeled advice-taking in terms of three components: accepting help, improving judgment, and sharing responsibility. For a recent review on advice taking and decision making, see Bonaccio and Dalal 2006.


Combine the model forecasts with the manager’s forecasts. Using 5 examples the authors show that a combined forecast produced by averaging the two forecast sources outperforms both the model and the manager. A 50-50 weighting proved close to optimal.


Interpreting advice is a key part of any forecasting or decision making task. Factors influencing the outcome include the characteristics of the decision-maker, the number of advisors and the interaction between advisor and decision-maker (forecaster). Application of research on advice taking to forecasting is still in its infancy.


Many forecasts are produced which are a combination of a statistical or model based forecast with the judgment of the forecaster. Considers the different forms of combination. Concludes with suggestions as to how such combinations could be improved, identifying also research gaps.
Adjusting a statistical forecast to capture excluded information is the most frequent method of forecasting in organizations. Examines the forecasts in four organizations, showing the circumstances where such adjustments are valuable (where the forecaster possesses positive information). Develops error models that overcome judgmental biases and lead to more accurate forecasts.

Unaided judgment is rarely effective. Shows how effective some simple heuristics are when compared to complex statistical alternatives. Reviews research that tests formal models of heuristic inference including in business, health, and legal areas. Ignoring part of the information can lead to more accurate judgments than weighting and adding all information.

Develops and tests a model of how participants respond to advice in a forecasting task. Forecasters give more weight to advice when they are less expert than their advisors, but still take insufficient account of advice. Those who are more expert than their advisors still take some account of advice.

Part of a program of research by the author on the software development task. Presents evidence that the judgmental estimates of the software development effort are more accurate than model-based forecasts. A discussion of the reasons for this is included in the same issue of the journal.

Combining judgment with statistical forecasts leads to improved accuracy. The results of this study are based on laboratory experiments using real data and an exponential-smoothing statistical forecast. Results also showed that the two forecasting methods are best combined automatically rather than through the use of judgment.

Extrapolative Forecasting

After judgmental approaches and often in combination with them, univariate methods represent the next most frequently used forecasting methods. Of those, the exponential smoothing methods associated with Holt and Winters are most commonly used. Autoregressive integrated moving average models, first introduced by Box and Jenkins are more complex mathematical structures that in their early stage of development required considerable skill to execute successfully. More recently, algorithms built into specialized forecasting software have eased the skill requirement. Similarly some specialized software such as the Hyndman R routines (see *Forecasting Software*) is built around the
state-space approach. Mostly, though, it has found a use in academic studies. Chatfield’s 2004 widely used textbook provides a comprehensive introduction to univariate forecasting methods.


*Exponential Smoothing*

Exponential smoothing as a forecasting method was first developed by Robert Brown in 1959. The method he gave his name to is no longer used. Two methods that followed in 1960 by Charles Holt and by Peter Winters, now generally referred to as Holt-Winters are widely used and are described in every forecasting textbook. Holt’s original report for the Office of Naval Research was never published and was reprinted in 2004 (Holt 2004). The method and its developments are well described in the two papers by Gardner 1985; 2006. Forecasts of trend using Holt-Winters are straight lines, which would be unrealistic at extended horizons, leading to the idea of damping the trend towards zero, proposed by Gardner and McKenzie 1985. Exponential smoothing is a family of methods. The breakthrough by Ord and his colleagues was to construct models based on exponential smoothing methods, thereby giving statistical properties to the forecasts from the Holt-Winters method. In a later paper, Koehler, et al. 2001 show the implications of particular modeling assumptions when constructing prediction intervals.


Koehler, A. B., R. D. Snyder, and J. K. Ord. 2001. "Forecasting models and prediction intervals for the multiplicative Holt-Winters method." *International Journal of Forecasting* no. 17 (2):269-286. Examines the state-space exponential smoothing model where, in contrast to the original formulation of exponential smoothing, it is possible to derive prediction intervals for the point forecasts. A broad class of model is described and prediction intervals derived. Also examines model-selection procedures for distinguishing between variants of the core exponential smoothing model.
Autoregressive Integrated Moving Average Models

George Box and Gwilym Jenkins popularized ARMA and ARIMA models in their book that first appeared in 1970 (Box, Jenkins and Reinsel 2008) and described the method widely referred to as the Box-Jenkins approach. As originally proposed, the method required considerable skill and judgment, but specialized forecasting software now automatically identifies and estimates the best model. Seasonality adds a layer of complication and Chatfield and Prothero 1973 provide an illustrative example. Outliers and other disruptions to a data series affect many estimation methods. Tsay’s 1988 description of procedures is based on the ARMA model. The early 1980s saw a couple of developments of standard ARMA models. Fractionally integrated ARMA processes, so-called “long memory” processes were introduced in 1980 by Granger and Joyeux. Smooth transition autoregressive (STAR) models appeared about the same time. Dijk, et al. 2002 provided a review of developments up to that time and Tong 2011 gives a recent retrospective view.

Updated version of a classic text. Emphasis on ARIMA models. Includes some more recent developments, e.g., GARCH models, unit roots, but also little-used methods such as transfer functions. Graduate-level text.

Walks the reader through the B-J procedure. Shows the importance of initial conditions: whether the initial values of the residuals are set to zero or are found by backcasting. Provides a good summary of issues with the B-J method and with forecasting generally.

Comprehensive. Covers basic STAR model and extensions, transition function choices, testing against linearity, misspecification tests (autocorrelated errors, remaining non-linearity, parameter constancy), modeling cycle (specification, estimation, testing). More emphasis on aspects such as model evaluation by means of out-of-sample forecasting and impulse response analysis, and the influence of possible outliers.

Original exposition of fractional ARIMA (ARFIMA). Fractional differencing corresponds to the expansion of \((1-B)^d\), \(d<1\). Generates a class of time series with low frequencies and provides potentially useful long-memory forecasting properties. Shows generation and estimation of these models.

A review with a long view. About 85 references. Fairly technical.
Detailed description of additive and innovational outliers, permanent and transient level shifts and variance change. Procedures illustrated on three real data series. Unfortunately, forecasts are made for just one set and outlier removal does not improve forecast accuracy.

**State-space Models**
The state-space approach is of general applicability but has most often been applied to the estimation or forecasting of a single variable, the dependent variable in the observation equation. Components on the right-hand side of the observation equation are the dependent variables in the state equations. Disturbances on the equations are parameters to be estimated by the filtering algorithm proposed by Kalman in 1960, a highly-cited though probably little-read paper intended for engineers, not forecasters. For a more accessible description of Kalman filtering see Meinhold and Singpurwalla 1983 and for a more rigorous treatment from a Bayesian perspective, Harrison and Stevens 1976. Harvey 1984 brought state-space modeling to the attention of time-series analysts and forecasters. Estimation of these models is demanding and Young, et al. 1999 showed how this can be done in the frequency domain. Durbin and Koopman 2012 give an up-to-date treatment. Hyndman, et al. 2002, 2008 propose a new state space formulation suitable for characterizing many time series methods and show their performance to be comparable to other established methods.

Follows the spirit of Box and Jenkins’ atheoretical correlation-based modeling. Chapter 1 provides the basic state space model and background to linear and non-linear as well as non-Gaussian models. Chapters 2-9 deal with the linear Gaussian state-space model, chapters 10-14 with a multiplicative nonlinear model. Graduate student level.

Introduction to Bayesian methods of forecasting and use of the Kalman filter. Shows that distinction between this and a standard state space formulation lies in the use of priors and estimation of parameters and hyperparameters. Insights into some applications. A stimulating discussion is also included.

Expresses exponential smoothing, (additive) Holt-Winters and ARIMA models as special cases of a state space model. Illustrates use of the Kalman filter.

Explains their state space approach to a number of different forecasting methods, in particular exponential smoothing. Also includes non-linear models, ARIMA, multiple seasonality as well as various applications.

Introduces the single source of error state space representation of a wide range of exponential smoothing models, discusses how they can be identified and estimated and concludes with empirical comparisons on the M and M3 data sets. The results show comparable performance particularly for short horizon seasonal series.


Brief article. Relates Kalman filtering to Bayes theorem and Bayesian analysis using terminology familiar to statisticians.


Young, an engineer by background, and colleagues examine the state space formulation for a seasonal unobserved component model, but argue that estimation in the frequency domain is superior to the standard maximum likelihood methods, both computationally and for faster convergence.

**Econometric Forecasting**

Econometric forecasting is a branch of econometrics. Unsurprisingly, while there are many general books on econometrics, there are few that specialize in forecasting. Those that do will include chapters on unit-root testing, vector autoregression and cointegration. More than other forecasters, econometricians seem inclined to self-reflection, perhaps because the method is complex, expanding and gives rise to disagreements as to the ‘best’ preferred approach. Econometric forecasting has also been less successful than its proponents would have wished.

**Econometric Forecasting Books**

Books listed here might be called books on forecasting by econometricians. Most take a broader view of forecasting beyond just econometrics. They require a fairly high level of statistical skill and familiarity with econometric methods. Hendry set himself the task of developing a theory of forecasting and two books are the result (Clements and Hendry 1998, 1999). They cover linear and particular non-linear multivariate models, respectively. Enders 2009 is a rigorous treatment of multivariate forecasting models excluding state space. Turn to Maddala and Kim 1999 for the reference source for unit-root and cointegration test methods extant in 1999.


A theory of forecasting: formal analysis of the models, procedures and measures of economic forecasting accuracy. Contains a taxonomy of forecast errors. Style and use of real and artificial examples representative of the Hendry approach to modeling and analysis.
Continuation of their first book, extends theory of forecasting to structural breaks in constant-parameter models. Investigates differencing, regime switching, intercept corrections as ways of dealing with forecast failure.

Graduate level textbook. Early chapters cover mathematical techniques. Followed by volatility analysis (ARCH and GARCH), trends and unit root testing, multiequation analysis (VAR and cointegration), concluding with non-linear modeling and testing.

An encyclopedic review of unit root and cointegration testing up until the late 1990s, presented in a mainly non-rigorous way. For researchers and graduate students.

Commentary on the State of the Art
There are a number of commentaries on how econometric forecasting has developed, for example, Diebold 1998 see *Macroeconomic forecasting*, and some commentaries on econometrics in general, for example, Pagan’s 1987 comparison of the methodologies associated with Hendry (the LSE approach), Leamer (an approach that is Bayesian in spirit) and Sims (vector autoregression). While Hendry’s and Sims’s approaches are frequently applied to forecasting problems, there are few assessments that bear directly on forecasting. Hendry’s 1980 article is an edited version of his inaugural public lecture. It is for the most part highly readable and intended for a lay audience. Appreciation of its finer points, however, requires more than passing acquaintance with econometrics and with research. Hendry provides a lengthy list of econometricians’ problems. Some are capable of diagnosis and in today’s best practice are routinely tested. Improvements in such practice are probably due in large part to his exhortation to “test, test, test,” which appears toward the end of the paper. The problems that existed 30 years ago (and for decades before that) still seem to be present today. It is worth remembering that misspecification tests (usually on residuals) point to problems in a model. They rarely provide a path to a solution. Zellner 1979 proposed a combined Bayesian and structural approach to overcome the problems of poor forecasting. Granger 1996 asked if we can improve the perceived quality of forecasts. His first request was to provide prediction intervals, preferably 50% limits, not just point forecasts. Today, many years after he wrote this, prediction intervals are still not commonly provided as part of the forecasting output. Where distributions can be calculated, by collecting the point forecasts from supposedly independent forecasters, he observed that they have a tendency to underestimate actual change and so provide an indicator of a structural break. When so indicated, he suggested a switch from structural models to adaptive ones until enough information accumulates to re-estimate parameters. This proposal does not appear to have been adopted. He also suggested paying closer attention to whether forecasts are conditional or unconditional and if the former, the policy assumptions on which they are based.
Calls for more emphasis on forecast distributions, not just point forecasts, a recommendation that is still relevant. Also asks at what horizon a forecast has value (or where the forecaster displays some skill), a topic that has since generated a small literature.

Readable and intended for a lay audience. Shows a spurious regression. Includes Hendry’s best known dictum: “test, test, test.” Notes the poor quality of economic data and how converting from nominal to real values can change the relation between income and expenditure.

Gives a detailed, critical step-by-step description of the methodologies associated with Hendry, Leamer and Sims. Avoids taking sides and concludes that none is a complete solution and econometric forecasters can benefit by using features of each approach.

Zellner, a Bayesian statistician introduces the SEMTSA (Structural Econometric Modeling Time Series Analysis) approach to economic modelling, contrasting this with classical econometrics. He argues if this is rigorously adopted it would lead to improved models with better forecasts.

*Unit Root Testing*
The implication of a unit root in a time series was not widely appreciated until the path breaking work of Box and Jenkins that first appeared in 1970. (See Box, Jenkins and Reinsel, 2008 in * Autoregressive Integrated Moving Average*). Essentially the problem occurs when time series contain stochastic trends. Following the original exposition of testing for a unit root (Dickey and Fuller, 1979), an explosion of articles reported on results of the test, on problems with it and alternative tests. One reason unit-root testing became so popular is the article by Granger and Newbold 1974 on spurious regressions. (See under *Cointegration *): when one random walk variable was regressed on another the expression had a high $R^2$. The nature of the problem was revealed by the low Durbin-Watson statistic. Although the original DF test is no longer used, because in general the residuals will not be white noise, the Augmented Dickey-Fuller (ADF) test is still the most commonly reported unit-root test. Both tests require non-standard t-tables that are built in to many software packages. The most popular of the remaining unit-root tests are Kwiatkowski, Phillips, Schmidt and Shin (KPSS, 1992) that starts from the null of no unit root, and the DF-GLS test (Elliott, Rothenberg and Stock, 1996) that does not require the complex strategy of the DF test. Using the DF test requires a complex strategy, making assumptions about whether or not deterministic trend or a constant or both are part of the data generating function. A recent paper by Harvey, Leybourne and Taylor 2009 suggested another simpler strategy (which they call “union of rejections”) that relies on using demeaned or detrended variables in the original form of the ADF equation shown above, and even better is to use this approach and the DF-GLS test. Nelson and Plosser 1982 in probably the best-known and most highly cited application of unit-root testing,
concluded that of 14 long macroeconomic series all except unemployment rate contain a unit root, that is, are best described as difference stationary rather than trend stationary. Their conclusions were surprising and their data have been re-examined many times. Various authors have extended their analysis to include structural breaks. Narayan and Popp 2010 cite the most important and compare results of the different studies, including their own. Much more important to forecasters, Franses and Kleibergen 1996 compared forecasting performance for the original 14 series both imposing and not imposing a unit root and concluded that imposing a unit root is preferred. Their work and limited evidence elsewhere establishes the forecasting principle on unit roots: when in doubt, impose them.

The original exposition of the unit-root test that bears their name. Provides tables of critical values of the test.

Proposes a more powerful test than the standard Dickey-Fuller test by first performing a generalized least squares regression of the series on a constant and trend then subjecting the residuals to a standard augmented Dickey-Fuller test without constant or trend. Provides a table of critical values and tables demonstrating improved power.

For the 14 variables tested by Nelson and Plosser, for both one-step-ahead and multi-step forecasts, a difference-stationary model is generally more accurate than a trend-stationary model, often significantly so. Suggests that for such series it is more useful to impose unit roots than to test for them.

Simpler strategy (called “union of rejections”) for unit-root testing. Uses demeaned (that is, the residuals from a regression of the series on a constant) or detrended (that is, the residuals from the regression of the series on a constant and time) variables in either the ADF or DF-GLS test of Elliott, Rothenberg, and Stock.

An alternative to the Dickey-Fuller test that starts with the null hypothesis of stationarity against the alternative of a unit root. Has not gained widespread use. Its use is illustrated on the Nelson-Plosser data, reaching different conclusions from applying the DF test.

One of many articles proposing a new unit-root test based on the Dickey-Fuller test. It provides, in table 9, a comparison of results of the Nelson-Plosser data by the original authors and seven other studies including the present one.

Nelson, C.R. and C.I. Plosser, C.I. 1982. “Trends and random-walks in macroeconomic time-series - some evidence and implications.” *Journal of Monetary Economics*, 10, 139-162. Examined 14 long annual macroeconomic series and, except for concluding that one (unemployment rate) is stationary, found that all contain a unit root, that is, are best described as difference stationary rather than trend stationary. The series covered spans of from 60 to 110 years, long enough for deterministic trends to be evident.

Vector Autoregressions
At about the time that unit-root testing was becoming popular, classical simultaneous equations models of the economy were not doing a good forecasting job. Although the multivariate extension of an AR model seems like an obvious alternative, it was not until Sims 1980 popularized vector autoregression that this approach gained much attention. Sims complained about the “incredible” identification restrictions used in classical structural simultaneous equations models. He noted that reduced form models can be identified more reasonably and, despite objections found in many textbooks, can be used in policy analysis. They are preferred for forecasting. Even simple VARs are “profligate in parameters” and Litterman 1986 described what has become a standard Bayesian approach to imposing parameter restrictions. The original VAR is a reduced form model and provides no information on what causes forecast changes, nor on what impact policy shocks would have on the economy. For these structural VARS are needed and various methods have been proposed (parameter restrictions, sign restrictions) to identify the structural model of the economy. Stock and Watson 2001 contrast the various roles of macroeconomic models.


Sims, C. A. 1980. “Macroeconomics and reality.” *Econometrica* 48, 1-48. The original reference on vector autoregression. Although he demonstrates the use of VAR in forecasting, all forecasts in the paper are actually within-sample. Notes in passing that shrinkage estimation methods such as Bayesian would be helpful.


Cointegration
Regression of two totally unrelated non-stationary time series, i.e., a spurious regression, can (and usually will) result in a high R-squared and impressive t-statistics. Yule’s 1926 paper is regarded as the
original demonstration. The standard reference to spurious regressions is the paper with the same title by Granger and Newbold 1974. They also illustrated that a test on autocorrelation in the residuals of the regression will detect the existence of such spuriousness. But what if a residuals test fails to find a problem? Sargan, in an inaccessible 1964 paper is credited with the creation of an error-correction model (without using the term). More than two decades later, in the most highly cited paper on time series, and also cited in their 2003 Nobel Prize award, Engle and Granger 1987 provided the complete answer. First we need non-stationary variables as discovered by the then relatively new unit-root tests. Then we test the residuals from ordinary least squares regression on a set of variables (without lags). If this single-equation residuals test shows no unit root then the variables are cointegrated, and there are implied parameter restrictions on the initial general model which can be imposed by inserting the residuals into an equation with variables in first differences – an error-correction model (ECM). One of the earliest applications of the ECM is the paper by Davidson, et al. 1978 though it did not actually use the term. Hendry now prefers the label “equilibrium-correction model”. Granger’s 1981 article is the original connection between error-correction models and cointegration, though expressed in spectral-analysis terminology. The Engle-Granger and similar single-equation residual-based tests have two problems. First, they test for only a single cointegrating vector against the alternative of no cointegration. Second, in using the ADF test on the residuals of the single equation they impose restrictions on parameters, limiting the equation dynamics. Johansen 1988 proposed a multivariate extension of the ADF that overcame these problems. Johansen and Juselius 1990 described the process and illustrated it with two examples. Johansen’s method has found its way into several software packages, which also incorporate the critical values of the likelihood-ratio tests for the restrictions (the maximum eigenvalue and trace tests). Although Johansen’s methodology is typically used in a setting where all variables in the system are I(1) (i.e. stationary after first differences are taken), having stationary variables in the system is theoretically not an issue but the presence of I(0) variables will result in a reduced rank matrix (i.e., indicate the presence of cointegration). Useful information can be gained from unit-root pretesting before embarking on cointegration analysis. The same issues of appropriate lag length, inclusion of deterministic elements, and problems with near-unit-root variables arise in each case.

Shows the early ideas of Hendry at work as he developed the LSE (London School of Economics) approach. Takes three earlier studies and with some modification to each nests them in a more general model. Introduces the notion of encompassing. Introduces (without naming it) the error-correction model. Pioneers the use of the test of out-of-sample forecast performance.

seven test statistics, the augmented Dickey-Fuller test on residuals of the ECM is recommended and most commonly used.

The standard source to demonstrate the danger of regressing one time-series variable on another and the need to examine the residuals for autocorrelation.

Original exposition of the multivariate equivalent of the Dickey-Fuller unit-root test usually referred to as the Johansen approach. Highly cited but less accessible than Johansen and Juselius.

More readable explanation of the method of determining the number of cointegrating vectors in a system than Johansen’s original article, with an example.

Non-linear Computer-intensive Methods

Non-linear approaches are computer intensive. They can be divided into model-based and data-driven approaches. De Gooijer and Kumar 1992 surveyed the first group including bilinear, SETAR (self-exciting threshold autoregressive), STAR (smooth threshold (or transition) autoregressive) and ARCH (autoregressive conditional heteroscedasticity) models. All non-linear models offer the forecaster considerable flexibility but at the expense of having to choose between a large number of alternatives. Overall, their evaluation suggested that there was little empirical evidence to show that non-linear statistical models outperformed simpler alternatives, at least with any regularity. An alternative approach without an explicit statistical model, Neural Network methods, sprang out of research in the computer- science community based on novel non-linear optimization algorithms (Rumelhart and McClelland 1986). Whilst their basis is an analogy with the functioning of the brain, in a forecasting setting they offer a method for establishing a non-linear relationship between a set of inputs and output(s). Mostly they have been used in predictive classification where objects (customers, companies, etc.) are classified into two or more categories, for example, in credit scoring and customer relationship management. This is an example where cross-sectional observations on objects of interest (such as consumers) are collected more or less at the same time (such as over a particular month). These computer-intensive methods have been subject to major criticisms, in essence because there is no agreed methodological approach to model building. Time-series-forecasters doubted that these methods would outperform more established approaches, and articles by Adya and Collopy 1998 and Zhang, et al. 1998 tended to support the skepticism. Nevertheless recent evidence suggests that such non-linear methods can improve forecasting accuracy. Since the 1980s, research proposing new methods and examining aspects of their performance has grown rapidly. Crone, et al. 2011 recently conducted an extensive competition. Again, results were mixed. Surprisingly, given what are regarded as
minimal data requirements, some of the neural network models performed well on short series. In one key application area - electricity load forecasting - neural network methods have generated a lot of interest and are believed to outperform alternative methods, see *Electricity and Call Center Forecasting*. Teräsvirta, et al. 2005 compared neural nets with STAR and linear models finding somewhat tentatively that the STAR models have the performance edge. Tay 2001 used a new method, support vector machines, an extension of NN concepts which he applied to financial time series. Research on using a hybrid approach of a neural network and a statistical model (usually autoregressive) dates from about the mid-1990s. Khashei and Bijari 2010 are a recent example.

Evaluated 48 neural network studies. Found 27 effectively validated (in comparison to established models, with out-of-sample forecasts and, sufficient sample size). Of these, 11 effectively implemented (ability to learn, ability to generalize out of sample, stability on different data), 11 did not but had better forecast performance than comparative models. Overall, 19 of the 22 contenders out-performed comparative models.

Describes the operation and results of a forecasting competition that included 22 neural network models, 5 advanced statistical models and 12 benchmark univariate methods on 111 monthly series. Results are mixed. A combination of the 5 best neural network models is consistently good, but individual models are often worse than simple benchmarks.

Surveys a range of autoregression-based non-linear models (including SETAR, STAR, ARCH). Notes limitations of tests for various non-linearities, the computational burden of making multi-step forecasts, and that their performance in comparison with linear-model forecasts is unclear.

Develops novel hybrid ANN-ARIMA model that outperforms either alone and Zhang’s 2003 hybrid model on three sets of real data. Lists 75 references.

The first full exposition of connectionism, the basis for artificial neural networks.

Introduces the use of support vector machines to business forecasting problems using financial time series. A comparison with back propagation neural networks shows them to be superior but sensitive to choice of user-selected parameters, a clear danger.


Compares linear AR models, STAR models and neural networks. Warns that specifying the non-linear models must be done “with care”. They found the STAR models slightly outperform NN with the linear models worst.


Extensive literature review. Describes the ANN paradigm; application areas, notably financial and electric-load forecasting; modeling issues: architecture of the system, selection of the activation function (usually logistic); training algorithm (usually backpropagation, a steepest-descent method); and distribution between training and test sets of data. ANN often no better than statistical methods in 24 studies surveyed.

### Comparing and Combining Forecasts

Forecasting competitions came into vogue in the 1970s. At their heart they are “horseraces” using a large number of real data series and intended to discover the best forecasting method out of those examined. While some poor methods were weeded out, the goal of forecasting competitions has still not been realized, though we have some understanding of the reasons for the observed differences in performance. When comparing a method against others over many series, one key issue is what statistic should be used to measure the forecast errors. Scale dependent measures, such as mean squared error, will overemphasize errors from numerically large series and should not be used in comparisons across several series. Even after using a suitable error measure another question remains: are the differences in rankings real differences or simply due to chance. Methods described in *Comparison of Forecast Errors* are intended to compare forecasts, whether produced by a model, from a survey, or by other means. They are not intended to help select the best model, using the term in its broadest sense and not necessarily a formal model. Testing for model specification is not the same as comparing forecasts. One surprising result from forecasting competitions was that a combination of forecasts, typically just a simple average of three or four forecasts, was often more accurate than the best individual forecast. Encompassing asks the question whether one model captures any information not already included in another model and a number of tests have been devised to answer that question. Combining and encompassing are different sides of the same coin. If one model does not encompass another then a composite forecast that averages forecasts from the two models will in general be more accurate than either individual forecast.
Forecasting Competitions

A crucial issue that came to prominence in the 1970s was the question of which particular method of forecasting was best and under what circumstances. Such comparisons, the essence of science, had been little studied until Reid’s analysis published in 1972 examined a variety of different univariate extrapolative forecasting methods. He attempted to explain the relative performance of forecasting methods in terms of the characteristics of the data series such as seasonality and stability of its trend. Newbold and Granger’s 1974 follow-up study is accessible and much better known. Both presentations included audience commentary which gave a flavor of the controversy such competitive comparisons generate. A number of competitions followed, including Makridakis and Hibon 1979, the precursor to the most influential competition of all, often referred to as the M-competition (Makridakis, et al. 1982). The latter used more data series (1001), more methods, and quite crucially, in order to overcome the critical comments directed towards Makridakis and Hibon 1979, the forecasts from the more advanced methods were carried out by independent experts. Further competitions followed including a real-time competition, M2 (Makridakis, et al. 1993), and the M3 competition (Makridakis and Hibon 2000) the largest yet, with 3003 time series analyzed. Crone, et al. 2011 (in *Non-linear Computer Intensive Methods*) have extended the comparisons to include a variety of neural network methods. Some conclusions of the competitions were surprising, but have stood the test of time: simpler, more constrained methods tend to outperform heavily parameterized, complex ones, combining forecasts from different methods beats most single methods, and there is no overall best forecasting method (Fildes, et al., 1998). While some methods have been shown to be generally inferior performers, what has been learned is that the method that forecasts best in the past is unlikely to be the best in the future and that variation in the performance of an individual method is considerable. Key issues are the range of methods, the type of data to which they are applied and how the resulting accuracy measures are summarized (Tashman 2000). The idea of finding a model selection algorithm that would recommend a particular method to the forecaster, once the characteristics of the data are established, remains an ideal that subsequent researchers have sought with only limited success. How has the statistics community responded to these findings, which cast doubt on the standard ARIMA models used by most academic researchers in business and economics? Fildes and Makridakis 1995 through a citation analysis showed that time-series statisticians had little interest in incorporating these empirical findings into their modelling. As in most areas of science, new findings have a difficult time overcoming the received wisdom of the established paradigm.

Simple model specifications often outperform complex alternatives. More general methods will not typically outperform constrained alternatives. Combining forecasts generally improves accuracy. Methods tailored to the specific characteristics of the time series under analysis will outperform benchmark methods. Concluded that how well a model fits the data has limited if any consequences for its forecasting performance.

Summarizes the major forecasting competitions. Notes problems with usual statistical assumptions: constancy or stationarity, use of logarithmic or other transforms, expected advantage of more complex
models, underprediction of forecast distributions. Shows by citation analysis that the competitions have been largely ignored by theoretical statisticians.

Compresses forecasts for 1001 series using 15 core methods and more accuracy measures than previous studies. Series were segmented into various subdivisions in search of an explanation as to the circumstances in which one method outperformed the remainder. Difficult to extract findings from the large number of tables.

Five forecasters received 29 actual series (23 company data, 6 macroeconomic) and made monthly forecasts up to 15 months ahead. Repeated one year later. Participants could use any additional available information, avoiding a major criticism of the M-Competition. Results show few or no differences in post-sample forecasting accuracy when compared with previous competitions.

Compares the accuracy of forecasts for 111 series using 13 core methods including variations of exponential smoothing, adaptive smoothing and naive benchmarks based on the random walk, plus combinations. Finding replicated in subsequent competitions.

Extends previous M-competitions: 24 univariate methods (including commercial software, decomposition, expert systems, neural networks) on 3003 series. Reaches same conclusions: value of simpler methods, combining, importance of accuracy measure. Damped trend smoothing is on average the most accurate extrapolative forecasting method over heterogeneous data.

Compared the accuracy of forecasts for 106 series using ARIMA, Holt-Winters, stepwise regression and various combination methods on 106 series.

Concerned with how forecasting competitions should be conducted. Issues include out-of-sample testing, the importance of rolling origins, choice of error measures and choice of sample series.
Measurement of Forecast Accuracy

There has long been controversy as to how forecast accuracy should be measured and forecasting methods compared. All measures of accuracy are based on a comparison between the realized actual (A) and the forecast (F), typically, A - F. The forecasting competitions showed that ordering methods by accuracy could change when different accuracy measures were used (See *Forecasting Competitions*.) Theil 1966 introduced a relative accuracy measure (Theil’s U2), still used today for comparing forecasts. Perhaps the most established measure when forecasting a quantitative time series is the root mean squared error where the measure is calculated over the available actuals. While it can be used to compare methods that forecast the same series, it should not be used for comparisons across series or involving several series. The early 1990s saw a series of papers arguing for different error measures, especially when comparing methods used on a group of series. Armstrong and Collopy 1992 and Fildes 1992 in two papers followed by a discussion took the approach of establishing criteria for what makes a useful error measure and then evaluated alternative measures against these criteria. They argue for robust measures, which excludes RMSE, and for the virtues of relative error measures such as:

\[ RAE = \frac{|A - F|}{|A - F_{Naive}|} \]

where the error from one method is compared with a naive benchmark forecast (such as the last actual, i.e., \( F_{Naive} = A_{just\ observed} \)). The controversy continued to simmer, however, with Clements and Hendry 1993, again with a discussion, arguing for a squared error measure when comparing across a number of related time series. In a recent contribution to the debate, Hyndman and Koehler 2006 develop a new method, suitable for comparing across series and, they claim, robust to some of the problems encountered when using RMSE, MAPE or RAE. Gneiting 2011 provides a recent review and recommendation that brings together both past academic work and surveys of practice. Surveys as summarized in Gneiting consistently find that practitioners favor absolute error measures over squared measures. An important problem arises when the data is categorical such as a weather forecast (of the probability of snow) or prediction of the probability of a recession. Murphy and Winkler 1992 proposed a measure for such problems. The economic consequences of forecast error and the links to standard accuracy measures were explored by Granger and Pesaran 2000.


Recommend Median RAE (MdRAE) for selecting the most accurate methods when few series are available and Median Absolute Percentage Error (MdAPE) otherwise. For calibrating a model recommend Geometric Mean of the Relative Absolute Error (GMRAE) (compares the absolute error of a given method to that from the random walk forecast).


Propose the generalized forecast error second moment (GFESM) as an invariant measure of forecast error. Twelve leading econometricians commented on the article, noting the measure’s limitations.

For choosing a forecasting method to be used on a set of similar time series, recommends the geometric root mean squared error. It is well-behaved and has a straightforward interpretation. The importance of evaluations using rolling origins is also demonstrated.

Argues that effective point forecasting depends on “guidance” or “directives,” either by disclosing the scoring function of the forecaster’s predictive distribution, ex ante to the forecaster, or by requesting a specific functional such as the mean or a quantile. Extensive citations and comparisons of standard accuracy measures, and rigorous theory.

Making a forecast is itself a particular form of a decision with economic consequences. This paper explores the link between statistical measures of accuracy and the economic consequences in a decision theoretical framework.

Propose a new measure the Mean Absolute Scaled Error (MASE) that avoids problems of most other measures. (A scaled error is error divided by the within-sample mean absolute error of the random walk.) Defines, compares and critiques almost all standard error measures.

Comprehensive description of measures for probabilistic forecasts including quantitative measures describing various aspects (or attributes) of forecast quality, including calibration (or reliability), refinement, resolution, discrimination, accuracy, bias, and skill.

Introduces the relative measure of forecast accuracy, known as Theil’s U2 to avoid confusion with his earlier measure, which should not be used to assess forecast accuracy.

**Comparison of Forecast Errors**
In forecasting competitions, regardless of the accuracy measure used, simple rankings do not tell the whole story. An important question is whether the differences are real or due to chance. Diebold and Mariano 1995 noted the difficulties in performing a formal statistical test. They observed that existing parametric tests depend on maintaining several assumptions about forecast errors (such as no serial correlation) that are unlikely to be met in practice. They proposed a test that is asymptotically distributed standard normal under the null hypothesis. The serial correlation correction is computed as a weighted sum of sample autocovariances. Through Monte Carlo analysis, Diebold and Mariano compared the size under both normal and fat-tailed distributions of their test, two standard non-
parametric tests and three parametric tests. As the tables of results reveal, the preferred test depends on sample size, though the Diebold-Mariano test is close to being correctly sized for samples larger than about 30. Harvey, et al. 1997 improved on the original. In a recent paper, Diebold 2012 bemoans the common practice of using the Diebold-Mariano test to compare models with a view to selecting the best. He notes that much of the substantial literature on both nested and non-nested testing that followed his seminal paper concerns testing models. Harvey, et al. 1997 presented an improved version of the test. Fildes 1989 addressed a practical but rarely considered question. When you have to forecast a large number of similar series do you need to compare forecasts from two (or more) different methods series by series or can you just compare forecasts on sample aggregate? For practical purposes there seems little to lose by selecting the method best on the aggregate set of series.

Highly cited paper. A standard F test can determine whether two sets of forecasts that are normally distributed, serially uncorrelated and contemporaneously uncorrelated are significantly different. The test proposed here overcomes the problem when these assumptions break down, and the F and other previous tests are incorrectly sized.

Diebold complains that much of the literature uses DM-type tests for comparing models in (pseudo) out-of-sample situations. He emphasizes that the DM test was intended for comparing forecasts (including without knowledge of, or even existence of, a model). Sees limited use of out-of-sample tests for model comparison.

Uses geometric root mean squared error (GRMSE) to choose which of two univariate methods to use on a relatively homogeneous set of time series. One series at a time ("individual selection") potentially better than choosing overall best method ("aggregate selection") but not in practice especially at short lead times. Better performance through individual selection difficult to achieve.

Propose a modified Diebold-Mariano statistic that avoids the size distortion (too many rejections under the null) when forecasting two or more steps ahead. For small samples several steps ahead the test is still over-sized but less than the original DM test.

On Selecting the Best Model
Related to the question of whether differences in accuracy between methods are real or due to chance is the issue of specification search. Data snooping, that is, reusing the same set of data whether for model specification or inference, is acknowledged to be a dangerous practice and to be avoided. Its close relative, data mining, is regarded as a good technique. Both are inevitable; for time-series work
there is often only one series available; for large data sets some systematic search is preferable to an investigator’s personal whim. Ashley, Granger and Schmalensee 1983 solved the issue for a particular problem: whether a given variable should be included in a regression as a causal variable or not. Swanson and White 1997 re-estimated models in a rolling window set-up but do not change specifications, a fairly common practice to select the best of the models being considered. They used the Diebold-Mariano test to determine if differences in out-of-sample forecasts were significant, despite Diebold’s 2012 later objection to the reliance on out-of-sample information (*Comparison of Forecast Errors*). The research perhaps motivated White 2000 to propose a formal test of the hypothesis that the best model discovered by snooping is no better than the benchmark. While the best model out of many specifications tested appears to be significantly better than the benchmark, a test that recognizes the effects of repeated testing shows that it is not. Hansen 2005 developed an improved version of the test. David Hendry is well-known for his dictum “test, test, test” and perhaps unsurprisingly has been the recipient of data snooping criticism. He has made the point that the automated general-to-specific model search algorithm developed by his group (Hendry and Krolzig 2005) uses misspecification tests only once to determine the appropriateness of the initial model. The search algorithm uses test statistics to guard against invalid simplifications, but only a single decision is required to select the final model, and thus, it is argued, the approach is immune to the data snooping problem (Hendry and Krolzig 2004).


Although testing for Granger causality, the authors point out that the test of “X causes Y” is a comparison of out-of-sample forecasts of Y from two models, one including X and one not. Provides the test to select between models.


Proposes a test in the same framework as White’s 2000 reality check more powerful and less sensitive to the inclusion of poor and irrelevant alternatives. Answers whether a particular forecasting procedure (or “model” in its broadest sense) is outperformed by an alternative.


Points out that the general-to-specific strategy involves testing for mis-specification only once. The title is a sly dig at the 1997 American Economic Review article by Sala-i-Martin titled “I just ran two million regressions”.


Detailed description of the general-to-specific modeling package developed by Hendry and co-workers. Involves repeated testing but avoids the data-snooping accusation by using such tests only once for mis specification as opposed to search within a specification.

Compares forecasting performance of survey of professional forecasters, univariate, multivariate (VAR) and neural networks using rolling samples both with and without re-estimation for 9 macroeconomic variables. Several error measures used on out-of-sample 1 and 4 step-ahead forecasts and Diebold-Mariano test applied to determine if significantly different. Re-estimated VAR proved to be best.


Data mining is regarded as good but the similar data snooping (reusing the same data more than once for inference or model selection) is bad. Illustrates the reality check test. Theory quite technical. Best out of 3654 potential models not significantly better than simple benchmark, therefore cannot be used for model selection.

**Combining**

Combining, the averaging of several forecasts, has emerged (from forecasting competitions) as a strong strategy. It seems to work because each component forecast captures some aspect of the data. Like a portfolio of stocks, the aggregate avoids the extremes of any individual. Early analysts believed that weighting the combination towards the forecasts that were better in-sample would be the preferred approach. For example, Bates and Granger 1969, usually regarded as the seminal article on combining, emphasized approaches with variable weighting schemes. There were two surprises: the combined forecast was frequently considerably more accurate than the best of the individual methods, and for some of the approaches, the individual forecasts entered with negative weights. Overall, there was little difference in forecast accuracy among the various combining methods. In a follow-up, Granger and Ramanathan 1984 showed that earlier methods could be viewed as constrained regressions. The 1984 article proposed using unconstrained regression to estimate combining weights and demonstrated on one data set its superior forecasting performance compared with using constrained weights. Although the theoretical underpinnings support their empirical findings, other authors report opposite findings. Clemen 1989 provides a review of research on and applications of combining of forecasts including a selective annotated bibliography. While research on various fixed and adaptive weighting schemes are discussed, he reaches no conclusions about the preferred approach to combining, acknowledging that simple averaging is hard to beat. In aiming to produce practical guidelines for combining, De Menezes, et al. 2000 illustrated the complications. They considered simple averaging and six potentially adaptive methods of calculating the individual forecast weights. For example, regression-based methods can be re-estimated when diagnostic tests or tracking signals indicate the need. Stock and Watson 2004 considered a wide range of combinations of forecasts of growth for 7 countries, and again simple combinations outperformed more adaptive and selective weighting schemes. The use of an expert system weighting scheme provides an alternative approach (Collopy and Armstrong 1992). Overall simple combining has proved an effective strategy and is closely related to *Encompassing*, where the aim is to identify the contribution of the different forecasting models to accuracy. In recent years, interest has moved beyond point forecasts to combinations of forecast distributions. Timmermann 2006 addresses the issue of the success of simple averaging from a theoretical perspective. He also reviews empirical studies of different combining methods and touches on the more complicated question of how to combine forecasts when forecast distributions or quantiles are wanted.
Compared 5 different formulas for combining. Surprised that adaptive methods worked better than fixed weights.

Comprehensive review of theory and applications and suggestions for further research. Notes that simple averaging of forecasts is a robust and close to optimal weighting method.

Describes an expert system for combining forecasts. While the rules can be stated quite specifically, implementation of them requires judgment and substantial analyst time. Although such methods still hold promise, they have attracted little follow-up.

Compare 7 methods of combining and give practical guidelines for combining forecasts based on a minimum forecast error variance criterion. For similar error variances, use simple average. Otherwise use an optimal (weighted average), particular method depending on sample size.

Showed that conventional combining methods can be viewed as constrained regression (coefficients sum to one and without constant). But treated as an unrestricted least squares regression with an intercept, and if the individual forecasts are biased, the method will be superior to Bates and Granger's (1969) optimal method.

A full analysis of different forecast combinations compared to various benchmarks including dynamic factor models (see *Macroeconomic Forecasting – Current Issues*). The consistent performance of equal weights combining they see as arising from the “widespread instability in performance of individual forecasts”.

Emphasizes theory underlying forecast combinations (wider information set, differential effect of non-stationarities and structural breaks, misspecification bias). Whether to combine determined by encompassing test. Reviews empirical evidence of various combining techniques.
Encompassing

We are here concerned with forecast encompassing rather than model encompassing. Forecast encompassing addresses whether the forecast from an additional method adds any information to the original forecast. Essentially it is a model-selection procedure. Combining and encompassing are two sides of the same question (Diebold 1989). If the test of forecast encompassing concludes that neither model’s forecast encompasses the other, then a combination of the two forecasts is the logical solution. The idea of testing whether a forecast from a competing model adds any information to the forecast of the original model (i.e., whether the original forecast-encompasses the competitor or not) goes back at least to Nelson 1972. Chong and Hendry’s 1986 paper puts forecast comparisons in a broader context and begins by asking how to compare the performance of one multi-equation model with another. They note some of the problems and limitations of dynamic simulation and model encompassing as model-selection devices and conclude that outside-of-estimation-sample forecast comparisons can be an effective approach to model selection. Clark and McCracken 2001 have provided a new and better encompassing test while Fang 2003 demonstrated how the use of encompassing tests could be used to understand model mis-specification.

Chong, Yock Y., and David F. Hendry. 1986. "Econometric evaluation of linear macro-economic models." Review of Economic Studies 53, no. 4: 671-690. Judging models requires analysis of the system but dynamic simulation is a flawed approach. Proposes several tests including a new "limited information" test of forecast encompassing, which is derived and assessed. Test is based only on forecasts and requires no other data from a model's proprietors.

Clark, Todd E., and Michael W. McCracken. 2001. “Tests of equal forecast accuracy and encompassing for nested models.” Journal of Econometrics, 105, 85-110. Develop new forecast encompassing test. Provide critical values for it and two other encompassing tests. Monte Carlo experiments show post-sample tests are well-sized and new test is most powerful. For nested models ‘Equal MSE’ tests fail to reject the null that unemployment has no predictive content for inflation; each of the encompassing tests indicates that it does.


Fang, Y. 2003. "Forecasting combination and encompassing tests." International Journal of Forecasting no. 19 (1):87-94. Uses Fair and Shiller forecast encompassing test and a double-differenced variant on several univariate and multivariate models to show why competing forecasts may be combined to produce a composite forecast which is superior to the individual forecasts. Also uses the two tests to discover model misspecification.

One-quarter ahead out-of-sample forecasts of 14 endogenous variables in the FMP model compared with ARIMA and composite based on regression weights. Since for most variables ARIMA has significant weight in the regression, FMP does not encompass it.

**Forecasting Distributions**

To say something about risk requires estimating error variance, or better, the complete forecast error distribution. Work on distributions has a long and unnoticed history until about 1982. Chatfield 1993 reviewed much early work, favoring the term prediction interval. Though computationally, prediction interval and confidence interval calculations are identical, a confidence interval surrounds the estimate of a fixed but unknown parameter, whereas a prediction interval surrounds a random variable, the point forecast. Chatfield considered many ways of calculating prediction intervals, and which are preferable. He presented a catalogue of reasons why prediction intervals are too narrow, and how the problem might be overcome. As Tay and Wallis 2000 remarked, density forecasting in finance began in 1982. If ever an article could be said to have started a movement, a leading contender would be Engle’s 1982 paper on autoregressive conditional heteroscedasticity (ARCH). Engle’s student, Tim Bollerslev, proposed an autoregressive moving average generalization, GARCH (Bollerslev 1986). As Engle notes in his 2004 paper, it spawned “an alphabet soup of ARCH models”. These models permeate the financial forecasting literature (Engle 2004). Andersen, et al. 2003 examined the forecasting performance for continuously traded exchange rates of a vector autoregression (VAR) model with various assumptions about variance. Using forecast-encompassing tests (see *Encompassing*) the authors found that daily GARCH models of volatility are forecast encompassed by a VAR with realized volatility. But ARCH/GARCH models are not the only way to estimate and forecast variance. Poon and Granger 2003 presented a comprehensive review of 93 papers containing out-of-sample forecasts of volatility. In addition they included a discussion of time-series models for forecasting volatility, including a catalogue of important members of the ARCH family. To compare forecast errors from different volatility models, they described various methods including that proposed by Diebold and Mariano 1995 described in *Comparison of Forecast Errors*. The popularity of the ARCH family of models in financial forecasting did not necessarily translate into superior performance. As might be expected, their dominance has not gone unchallenged. Most importantly, use of squared daily return, the usual proxy for conditional variance, can result in selection of an inferior ARCH-type model (Hansen and Lunde 2006). Berkowitz 2001 described a new approach to the evaluation of density forecasts that made use of the entire distribution and could be used in samples as small as 100.


Continuous (or near continuous high-frequency intraday) data permit the more accurate measurement, modeling, and forecasting of daily and lower frequency return volatilities and return distributions. Demonstrated for the Deutschemark/Dollar and Yen/Dollar spot exchange rates., Vector autoregressive volatility forecast, coupled with a parametric lognormal-normal mixture distribution produces well-calibrated density forecasts of future returns.


35
Describes new approach to evaluating interval and density forecasts that examines entire distribution. Monte Carlo analysis shows that approach is well-sized and has better power than other common methods, even with samples as small as 100.

The generalization to an autoregressive moving average structure of Engle’s ARCH model. Developed by his student. Covers specification, estimation and testing.

Describes and compares use of theoretical formulas based on a fitted probability model and empirically based simulation and resampling procedures. Comments as to why prediction intervals tend to be too narrow in practice. Makes no mention of volatility forecasting and the ARCH family of models.

The highly cited article that started the volatility revolution and earned Engle the Nobel Prize. Suggests OLS regression and testing autocorrelation of squared residuals to see if ARCH model should be estimated. Illustrates with estimates of variances of UK inflation by maximum likelihood.

Edited version of his lecture on receiving the Nobel Prize. A readable history of ARCH and models derived from it with illustrations of their use in financial forecasting.

Realized volatility, a sum of squared intraday returns yields a biased estimate of volatility that worsens at higher sampling frequencies because of market microstructure noise. Noise may be ignored at low frequencies (e.g., 20-minute sampling) Show kernel-based estimators reduce bias at higher frequencies.

Of 66 papers comparing out-of-sample forecasts of volatility, 22 of 49 found historical volatility measures more accurate than GARCH models, 26 of 34 found option-implied standard deviation better than historical volatility and better than GARCH in 17 of 18 comparisons. Include catalogue of important members of the ARCH family.

**Forecasters and Forecasting Practice**

Forecasting as it is practiced in organizations is very different than what the researcher might infer from the technical articles referenced in other sections. Judgment is the most common method of forecasting
and it is a rare organization that uses even modestly complex econometric methods to support their forecasting. Perhaps only in governments do these get frequently used. Dalrymple 1987, a classic article, considered the issue of the adoptions of forecasting techniques and how well-used they were at the time. Other surveys tell similar stories and McCarthy, et al. 2006 summarize the results. They all tend to reach the same somewhat depressing conclusions. Despite the advances in the academic research literature informal simple approaches to forecasting continue to dominate organizational practice. Why is this? Is it because the methods themselves fail to work outside the hot-house research laboratory? Or is it because of limited resources in the firm? Few studies have examined this question. Fildes and Hastings 1994 studied forecasting in a multiple business unit multinational and found that resource constraints including the technical knowledge of the company’s forecasters constrained the performance of the forecasting function. Galbraith and Merrill 1996 examined the political issues that also limit the effectiveness of the forecasting function. But what should an ideal forecasting activity look like? Winklhofer, et al. 1996, Moon, et al. 2003, and the commentary in the same issue suggest some possibilities. An alternative explanation lies in the education, training and motivation of forecasters. The literature is extensive on such topics as: forecast bias, herding (where a forecaster aims to achieve a forecast close the consensus) and whether the forecaster in producing a point forecast is implicitly using a loss function that is not the usual squared error loss. Elliott, et al. 2005 is an illustrative example of the recent literature. Lamont 2002 discussed these issues in the context of macroeconomic forecasters. Ramnath, et al. 2008 (see *Accounting and Finance*) summarize the extensive research in this area with Hong and Kubik 2003 in *Accounting & Finance* providing an example. The results of the debate underline the fact that many factors motivate forecasters apart from the desire for accuracy and that unbiased and efficient forecasts will be the exception rather than the rule.

This survey of forecasting practice, now dated, gives an insight into the different factors such as firm size and industry that affect the use of different forecasting techniques. Judgmental techniques were generally preferred to quantitative alternatives.

Notion of a rational forecaster depends on the forecaster’s loss function;. Presents tests of rationality under a potentially asymmetric loss function. These then used to test OECD and IMF forecasts of government budget deficits, which are systematically overestimated, suggesting asymmetric loss. But are we surprised?

Surveys market forecasters in 10 divisions of a multinational. Describes in depth how forecasts are produced and the forecasters’ perceptions of inadequacies in the market forecasting process. Identifies barriers to effective implementation, in particular computer support and inadequate data. Concludes with recommendations as to how the forecasting activity could be organized more effectively.
Forecasting is often a political activity. Using a survey evidence from 64 companies, the authors demonstrate the damage done to forecasting accuracy and the quality of decision making by various types of political interventions in the forecasting process.

Macroeconomic forecasters as they grow older become more extreme with their forecasts and less accurate.

Mentzer and colleagues have been investigating forecasting practices in US companies for 20 years. This study summarizes and evaluates the changes observed over the period, contrasting the results with what is known of good forecasting practice. Actual organization of the forecasting function falls well below what could be achieved. The good news for researchers and consultants alike is there is plenty of scope for action research to improve performance.

Based on in-depth interviews in 36 US companies, this paper describes an ambitious attempt to categorize four stages of development of ‘best practice’ demand planning. Four dimensions were seen as valuable in capturing the diversity of practice: functional integration, methods used, systems and performance evaluation.


Applications

The range of management, business, government and economic applications of forecasting methods is wide. There are a limited number of areas where there have been substantial contributions both to the applications area but also to the development of novel methods. Operational (short term) forecasting occurs widely in most organizational settings with its focus on providing manpower and materials to support production, distribution and service. Energy demand, pricing and telephone call center demand have proved particularly important both in research and in practice. More aggregate forecasts of market demand are also needed for longer forecast horizons, both period by period and also the market potential; ideally the methods should include marketing instruments such as price and promotions. With
consumer durables such as mobile phones, new technologies with added features replace the established products and these hard problems again require new methods. The increased availability of consumer data has permitted work on a quite different and new type of problem, that of making disaggregate forecasts of individual purchasing patterns, with the objective of identifying profit opportunities and risks. Forecasting the adoption of new technologies has some parallels with forecasting sales of a new product, though it has developed methodologies all its own. It also has a specialized journal (see *Journals*). Forecasters should always ask the purpose of their efforts and one purpose is as input to planning and strategizing, where the limitations of the forecast need to be known. The first applications of forecasting methods were in macroeconomic forecasting and there is a large literature. Likewise, the accounting and finance area has been heavily researched and could support a bibliography on its own. Two key areas are forecasts of company performance (usually produced by industry analysts and models to forecast uncertainty Other application areas have a small and sometimes specialized literature. Tourism forecasting and sports event forecasting are two such areas.

**Operations**

Forecasting to support operations necessarily has a short-term focus with much of the environmental uncertainty regarded as fixed. Extrapolative methods predominate. Many applications such as retailing require forecasts of many series (often greater than 10,000 at high frequency, perhaps daily) so that automatic selection of an appropriate method is essential. Early on, Fildes and Beard 1992 identified some of the important issues facing operational forecasters in an organization. Forecasting many products has implications for the forecasting support system (FSS) used, the methods adopted and the error measures employed. Despite the importance of forecasting to operations, it has in practice generated relatively little research on the links between the two (Syntetos, et al. 2009). A key issue is how valuable improvements in forecasting are to an organization’s operations. Gardner 1990 provides an example of the trade-off between inventory level of spare parts and service. The data structure common in the demand for spare parts and in retail sales requires consideration of a quite different class of extrapolative model than those mostly commonly used. Intermittent demand occurs when there are periods of zero demand interspersed with spikes of non-zero activity. Smoothing methods produce forecasts of both the expected quantity and the expected time between the periods of non-zero demand. After more than 20 years of neglect following Croston’s original work, the topic is now seeing a lot of research activity. Syntetos and Boylan 2005 evaluated a number of different forecasting methods. Within a supply chain where a retailer places orders with a manufacturer (depending on expected demand, stock in hand and orders already placed) inaccurate forecasting at both retail and manufacturer produces potentially substantial losses. In particular, the retail ordering process leads to an amplification of the fluctuation in the manufacturer’s demand - the so-called bullwhip effect. Chen 2000 quantified the effects of the bullwhip. Despite the limited applicability of the models that have been examined, the problem of collaboration between retailer and manufacturer (Collaborative Planning, Forecasting and Replenishment or CPFR) is a topic of considerable interest to companies and software providers. Aviv 2001 presented a model of the value of shared information. After more than 10 years of a burgeoning and often contradictory research literature, the emerging conclusion is that additional shared information is usually valuable! Forecasting many hundreds of products, arranged within a hierarchy (such as stock-keeping unit, brand, category by day of the week, week, month etc.) has implications for the forecasting support system (FSS) used, the methods adopted and the error
measures employed. Athanasopoulos, et al. 2009 recently introduced new methods to utilize information in the hierarch to produce better top-level and bottom level forecasts.


Much of operational forecasting naturally falls into a hierarchy. An example considered here is tourism demand by country, region, sub-region and purpose-of-travel. Forecasts can be bottom-up from sub-region to country or top demand. Here a new approach is proposed which performs favorably compared with these two standards.


Collaborative forecasting (CPFR) potentially important for ensuring supply chain efficiency. Proposes a two-stage supply chain of retailer and manufacturer to explore the benefits of sharing information and in particular working with a common forecast. Useful for understanding key elements in collaboration – but the results depend on particular parameterisation. Shared information potentially valuable!


Develops a model of a two-stage supply chain of retailer and manufacturer. The orders experienced by the manufacturer fluctuate substantially more than demand placed on the retailer – the bullwhip effect. Shows how much sharing customer demand information across the supply chain mitigates the effect.


An early paper that covers many aspects of forecasting to support operations including A-B-C analysis, product hierarchies, error measures, monitoring schemes and forecasting support systems.


Presents a case study for the supply of spare parts in the US Navy. The Navy's standard forecasting method was exponential smoothing with a fixed smoothing parameter. Gardner evaluates his damped trend smoothing (see Gardner and McKenzie 1985, *Exponential Smoothing*). Using trade-off curves with his proposed forecasting method produces substantial improvements.


Evaluates different methods for forecasting intermittent demand including exponential smoothing, Croston's method and a new method proposed by the authors. Also considers the difficulties in using standard error measures.

Aims to synthesize three disparate literatures: system dynamics, control theory and forecasting theory (both statistical and judgemental). Statistical methods for inventory planning were neglected for many years. But control theory and system dynamics, in particular the concern with bullwhip effects have been more prolific. Concludes with inter-disciplinary agenda for further research.

**Electricity and Call Center forecasting**

Forecasting electricity demand and prices to support pricing and scheduling and forecasting in-coming calls into a center to schedule staff are operational problems. They have particular characteristics in common: high frequency data (half-hourly is typical), multiple seasonalities and a forecast horizon often of only one day. This has required operationalizing and evaluating newly emerging methods such as neural nets (Darbellay and Slama 2000; Hippert, et al. 2001), the development on new techniques to account for multiple seasonality (Taylor 2003), and the evaluation of a range of different methods including data segmentation for time-of-day- and, day-of-week effects (Conejo, et al. 2005; Taylor, et al. 2006). Weather is a particularly important variable to include, (Taylor and Buizza 2002). Telephone calls into a call center are count data and the models developed for them are based on Poisson processes which are time-of-day dependent. Multiple seasonalities are also potentially important. Gans, et al. 2003 surveyed the relatively small amount of research on call-center forecasting.


Compares a number of different methods including neural nets, ARIMA, wavelets, transfer function models and dynamic regression. Dynamic regression and transfer functions perform comparatively well. Naïve is surprisingly effective. This paper adopts techniques discussed in highly cited earlier papers by the same authors.


Much cited paper exploring issues in building neural network models for electricity load pointing out that to be successful requires significant non-linearity.


Describes the data properties and sources of uncertainty inherent in call center data. Reviews the literature on call-arrivals forecasting: use of descriptive models for investigating short-term patterns; explanatory statistical models incorporating time-of-day, day-of-week, holiday, variables; classical models based on queuing theory to produce point and distributional forecasts.

Reviews many papers using neural nets to forecast electricity load. The authors conclude that the cumulative evidence as to the improvements in accuracy achieved remains limited and flawed, although they are optimistic that something substantial underlies the hype.


Compared a variety of methods for forecasting electricity load including Neural networks with a regression based model of each hour’s load including temperature and lagged effects. Illustrates some of the hazards of evaluating load forecasting.


Compares a number of methods applied to two electricity demand series. Methods include double exponential smoothing to include the complex seasonality, ARIMA, neural nets and principal component regression. Exponential smoothing proved the best performer.


Uses a set of 51 weather predictions integrated into an ensemble. Permits a comparison between a load prediction (using a neural net) based on the point weather forecast and one averaged from the 51 weather scenarios and various benchmarks. Using weather scenarios produces more accurate forecasts for lead times of 1-10 days.

Marketing

Marketing models are usefully categorized into new product models where there is little directly relevant data and models for established markets. The models may forecast sales, market share, or market penetration. Armstrong and Brodie 1999 have an evidence-based non-technical review. The section focuses on new product models as they require distinctive methods from the standard econometric approaches that can be used when there are established markets with considerable data. Bass 1969 developed an adoption model applicable to a consumer or business product or service for which he derived an S-shaped logistic trend curve. is the curve describes an initial slow adoption, an increasingly rapid take-off, then a flattening of sales when most of the potential customers have adopted it. The problem with the method if it is to be used in forecasting is to how to estimate its parameters with limited or perhaps no data available (Goodwin, et al. 2013). Meade and Islam 2006 offer a wide-ranging review of trend-curve modelling for adoption. A contrasting approach, which aims to help in the design and marketing of new products, is to ask potential consumers about their intentions to purchase depending on the proposed characteristics of the new product or service. Their choices are then modelled, dependent on both the purchaser’s characteristics (e.g. income) and the product’s attributes, to produce predictions of the probability of purchase. The individual probabilities can then be aggregated to provide forecasts of the market as a whole. Urban, et al. 1996 took a simulation approach to modelling the uptake of a really new product, where there is no analogous product. The subjects of the experiment were asked to assess their probability of purchase (see Morwitz
A problem arises with standard models of consumer choice, when a new product is added that is similar to the other alternatives under consideration (Brownstone and Train 1999). Fader and Hardie 1996 are concerned with a more common marketing problem: the relationship between the SKU and the brand. Using choice models based on attributes that include different forms of the product e.g. different pack sizes, etc. they aim to forecast new-product extensions. This approach is in keeping with a research program on modelling and forecasting sales and market share data of established products. Appropriate methods are well covered in the book by Hanssens, et al. 2001.


Non-technical review of evidence to provide guidelines. Covers intentions, Delphi, role playing, conjoint analysis, judgmental bootstrapping, analogies, extrapolation, rule-based forecasting, expert systems, and econometric methods. Discuss which methods are most appropriate to forecast market size, actions of decision makers, market share, sales, and financial outcomes.


Formative highly influential paper describing the adoption of a new product (or service) Probability of adopting a new product depends on the number of people who have not yet adopted, their intrinsic probability of adoption, plus further impetus from those who have already adopted. Then predict the cumulative number of adopters at any point in time.


Models future demand for alternative fuel vehicles (car, van and truck: electric, methanol, natural gas). Standard assumptions concerning the alternatives and their independence (the IIA assumption) cannot hold. Develops a model that permits flexible substitution between fuels that depends on the type of transport. The results from the more flexible model gave quite different predictions than those from the standard approach.


Most marketing models based on brand choice. Develops a choice model using SKU attributes to include more managerially relevant forecasts to produce forecasts of brand-extensions. Illustrated by forecasting the effects of introducing a new product which is a ‘line extension’, that is it has similar attributes to those in the estimation sample.


Examines the value in using analogies in calibrating the parameters in new product diffusion models. The results are disappointing leaving practitioners facing high levels of uncertainty in their new product forecasting.

Gives a complete review of the research evidence and econometric issues in building market response models but is also firmly based on the practical issues facing the market forecaster.


Excellent highly cited review paper on diffusion of innovation and aggregate new product forecasting.


Describes a complex and expensive procedure that combine a number of core forecasting methods including simulating the buying decision in order to support the expensive decision to launch a new electric car, so any method that gives both insight and improved accuracy is of considerable value.

Technology forecasting

Technology forecasting aims to forecast the development and adoption of new technologies. Bass’s model of the adoption of a new product (see *Marketing*) was extended by Norton and Bass 1987 by including second and later generation technologies. They utilized the information in the adoption of earlier products (such as mainframe computers) to produce better forecasts of the adoption and market potential of later generations. Together with Bass’s original article these models of adoption of technologies or consumer durables have proved highly influential and led to a proliferation of models that attempt to take into account the particular characteristics of the market being modelled. Perhaps of greatest current potential is where choice-based approaches have been integrated into diffusion models (see for example Jun, et al., 2002). Martino 2003 has summarized recent methodological developments over the whole field of technological forecasting. Daim, et al. 2006 showed how data and text mining can be incorporated into a system dynamic model, an illustration of the novel methods developed in the field to deal with disparate data. However, few of them (apart from Delphi and trend-curve modelling) have been subjected to rigorous testing of their effectiveness.


Integrates forecasting tools such as scenario planning, growth curves and analogies with bibliometric and patent data, article’s most interesting feature. System dynamics is also used. Technologies being forecasted are fuel cell, food safety and optical storage technologies. Claims results helped to validate the proposed methods as appropriate tools to forecast emerging technologies.


Choice-based data on mobile telephony used to incorporate competition between providers and substitution between technologies into a diffusion model. The advantage of the approach is that it
embeds consumer choice into a dynamic forecasting framework. Produces more accurate forecasts than alternatives based on Norton and Bass.

Martino, J. P. 2003. "A review of selected recent advances in technological forecasting." Technological Forecasting and Social Change no. 70 (8):719-733. doi: 10.1016/s0040-1625(02)00375-x. Written by one of the most long-standing contributors to the area of technological forecasting, the paper offers an update on the most ‘recent advances’, most of which are established, such as growth curves and Delphi. Most useful for capturing the state of knowledge in technological forecasting; otherwise dated.


Planning and Strategy
The inadequacies of judgment (see *Judgmental Methods*) in strategy formulation and planning have often been documented with Hogarth and Makridakis 1981 presenting an early but still valuable review. But even when formal models are used their accuracy remains overly limited as far as many users are concerned. There inevitably remains considerable uncertainty and the authors propose novel approaches to planning which attempt to overcome an undue reliance on accuracy. Multi-attribute decision models can both help to structure complex decision problems and offer guidance as to where robust solutions might lie. In order to deal with the high levels of uncertainty and the interdependent nature of strategic decisions, Schoemaker 1991 proposed that scenarios offer a route forward. Scenarios, by which is meant “a consistent set of statements about possible future events and trends and their dependencies tracing the progression of the present to the future through a descriptive narrative” are not themselves forecasts. They are built using a variety of different forecasting methods, primarily structured judgment. What they aim to deliver is a means for an organization to both develop robust strategies and to understand the uncertainty it faces in making longer-term strategic decisions. Bunn and Salo 1993 argued the controversial view that scenario analysis has much in common with forecasting. That view seems correct, not least because scenarios are themselves composed in part of conventional forecasts. Whether we rely on conventional forecasting or scenario analysis, strategic surprises will occur. Makridakis and Taleb 2009 considered what can be expected of forecasting and how overconfidence in models can lead to ignoring features of the problem that might turn out to be all-important (as in the financial and economic crisis of 2007 on).


Review the extensive literature on judgment as it affects forecasting and planning, in particular the "illusion of control". Argue the balance should shift away from forecasting accuracy to identifying sources of uncertainty and the use of sensitivity analysis within a multi-attribute decision framework..


Argue that high levels of uncertainty continues to affect us, despite all the research on forecasting in the least 40 years, Critically, future uncertainty cannot be assessed. Serves as an introduction to other papers in the journal discussing these issues.


Defines scenarios. Discusses how scenarios should be constructed and used if they are to be most effective.

**Macroeconomic forecasting - overviews**

Macroeconomic forecasting and business cycle analysis and prediction were two of the first applications of formal forecasting techniques. They remain important to governments although perhaps of less interest to economists. In the 1930s, the early macroeconomic models particularly those associated with the Cowles Commission (renamed the Cowles Foundation in 1955) were influential in the development of econometric methods. But by the 1970s their performance was seen as disappointing. As with the *Forecasting Competitions* literature, when comparisons were made with much simpler models the results proved surprising; the large-scale system model was often out-performed by simpler autoregressive alternatives. Wallis 1989 surveyed both the history and the problems of making valid comparisons among forecasts from macroeconomic and time-series models, focused mainly on efforts in the United Kingdom. Diebold’s 1998 paper took both a longer and broader view, going back to the work of the Cowles Commission and examining in some detail the multivariate dynamics of vector autoregressions and related models. Although Fildes and Stekler 2002 focused on GDP and inflation forecasts for the UK and the US they provided an overview of issues in the appropriate use of revised series, measuring and comparing forecast errors, assessing whether one model forecasted better than another, whether that provides any guide to future performance, and whether irrationality (bias in forecasts, or inefficient use of all available information) indicates potential for improvement. Zarnowitz 1995, one of the major contributors to our understanding of the record of macroeconomic forecasters, examined results from individual macro forecasters (some of whom would use econometric or statistical models). He showed that individually many were biased and inefficient, particularly for the key variable, inflation. In responding to fluctuations in the macroeconomy it is particularly important to forecast the turning points in such variables as GDP. If we are bad at making point forecasts (which we are are) we are even worse at identifying turning points, even after they have happened. Geoffrey Moore made a major contribution over many years to this topic in a research program supported by the US National Bureau of Economic Research, with other distinguished economists including Arthur Burns, Milton Friedman and Wesley Mitchell. He co-edited a book that covered major aspects in the development and evaluation of leading indicators (Lahiri and Moore 1991). More recently, Stekler (2007) examined the
process of making macroeconomic forecasts, including models, data, the judgment of the forecaster and interactions among them, his aim being to identify potential directions for improvement. He concluded that judgmental or other adjustments to model forecasts should be documented and reasons for forecast errors investigated more carefully.

Examines the history of macro forecasting including non-structural VAR type models, arguing that the development of economic theory through dynamic stochastic general equilibrium models should lead to improvements in forecasting – but a lot has to be done before this is to be achieved. As yet no evidence has emerged to support this view.

Summarises the comparative accuracy of UK and US macroeconomic forecasts, presenting new evidence. Discusses comparative accuracy of macro models compared to time series alternatives, whether the forecasting record has improved over time, the rationality of macroeconomic forecasts and how to choose a forecasting service. Use of judgement unequivocally improves forecasts.

Twenty-two articles by international experts cover advances in three areas: the use of new developments in economic theory and time-series analysis to rationalize existing systems of indicators; more appropriate methods to evaluate the forecasting records of leading indicators, particularly of turning point probability; and the development of new indicators.

Macroeconomic forecasting accuracy has improved little from the 1970s. Stekler examines the evidence, identifies structural change as key and makes grounded suggestions as to how improvements could be achieved through a careful examination of how forecasts are produced and the information available to the forecaster.

Gives a brief history of UK macroeconomic forecasting and discusses how forecasts are produced through macro models, interpretation of forecast errors, comparison between models and compared to time series forecasting models. Wallis and colleagues subsequently produced a number of books that compared the performance of various UK models.

One of the early papers that examines the macroeconomic forecasting record of individual economists. Systematic errors appear for inflation in particular. A combined forecast performs well.
Macroeconomic forecasting – current issues

Researchers continue to propose new methods of macroeconomic forecasting, stimulated by new data sources and computational abilities. In an early and highly cited paper Meese and Rogoff 1983 looked at the out-of-sample performance of structural exchange-rate models and concluded that random walk forecasts were as accurate. The debate on whether the poor performance was the result of parameter instability or misspecification is still continuing. How to deal with revisions of macroeconomic series and their effect on accuracy poses challenges to researchers. Diebold and Rudebusch 1991 examined the value of using a composite indicator to forecast GDP and the consequences of using preliminary data in a real-time analysis. Unfortunately, performance deteriorates substantially. Recognition of the interconnectedness of the world economy does not always lead to models that incorporate out-of-country information. Arnold Zellner applied Bayesian methods to the problem of predicting international GDP growth rates with methods that combined information from the international data set (Min and Zellner 1993). Such international linkages are now seen as particularly important. Marcellino et al. 2003 attack the issue as one of disaggregation. They found aggregating country-specific (i.e., disaggregate) forecasts were more accurate than using an aggregate model. Although that finding is common, the question of when disaggregate data yield better forecasts is still unresolved. Also a continuing issue is how to deal with many potentially relevant data series with relatively small numbers of observations in each. Among their many contributions to macroeconomic forecasting, Stock and Watson 2002 illustrate one approach, Dynamic Factor Models. A developing research issue is that of ‘nowcasting’ where real time data up to the present time is used to estimate some macroeconomic data series such as current GDP (typically published quarterly) based on provisional information supplemented by information from more frequently available data series such as retail sales (published monthly). Giannone, et al. 2008 introduce a method to incorporate many monthly data series to improve GDP quarterly estimates.


Influential paper which takes into account the time of release of economic data. Common academic practice of using ‘final’ published data does not conform to the reality of the practicing forecaster who have only the latest cohort of data. Using real time data leads to a substantial deterioration in accuracy.


Much cited reference to a currently important topic of practical significance – how to formally incorporate a flow of data releases to improve the estimates of current GDP which only become available later.


Compares forecasts of real GDP, industrial production, inflation, unemployment from univariate autoregressions, vector autoregressions, single equation models that include Euro-wide and US aggregates, and large-model methods in which forecasts are based on estimates of common dynamic factors. Developing disaggregate forecasts improves the accuracy of the aggregate.

A random walk forecast is as accurate out-of-sample for 1 to 12 months ahead as univariate, VAR and structural exchange rate models, despite using realized values of future explanatory variables in the structural models. Uses RMSE, mean error and mean absolute error measures, but does not test for statistical differences.


Growth rates for 18 countries using both country and world variables, estimated with fixed and time varying parameters(TVP). Evaluates various methods for combining forecasts (see *Combining*). Including world variables helped accuracy. Using TVP models was beneficial. Attempting to decide which methods to include in the combination did not improve accuracy.


Hundreds of lagged economic indicators are simplified using principal components into a small number of diffusion indices (alternatively dynamic factors) and the dependent variable forecast as a function of the diffusion indicators. Results show major improvements over autoregressive benchmark models suggesting that such simple models can capture economic variability.

**Accounting and Finance**

These two broad areas have generated many forecasting-related papers. They include prediction of company financial variables, including stock prices, and the appraisal of a company’s future financial prospects. Altman’s 1968 classic article on company bankruptcy used discriminant analysis to predict the likelihood of bankruptcy. At its simplest this is a 0 (going concern) or 1 (bankrupt) classification problem. Wilson and Sharda 1994 used neural networks to compare their results with those of Altman. Research on forecasting of company earnings has concerned itself with two rather distinct questions: whether the earnings forecasts produced by analysts outperform statistical models, and whether the forecasts are efficient, that is they cannot be improved on using only publicly available information. O’Brien’s 1988 much-cited study, focusing on the benefits of timeliness, found that analysts’ forecasts were more accurate than time series forecasts. A different perspective is to examine what motivates analysts who may have other objectives than accuracy (Hong and Kubik 2000: see also *Forecasters and Forecasting Practice*). A recent survey is given by Rammath, et al. 2008. The third important question: whether share prices can be predicted better than those derived from efficient market forecasts, most particularly the random walk model that states that price changes are independent. In many cases, they are also assumed to be identically distributed but as discussed in *Forecasting Distributions*, ARCH/GARCH models have been developed to overcome this particular limiting assumption. Timmermann and Granger 2004 discussed the circumstances where we might expect improvements over the random walk in an efficient market. An earlier illustration is a paper by Pesaran and Timmermann 1995, the purpose of which “is to assess the economic significance of the predictability of U.S. stock returns, accounting for the forecasting uncertainty faced by investors”. Thomas 2000 surveyed another aspect of the financial services industry – the assessment of consumer credit risk with
a view to granting loans to potential customers, a large-scale commercial activity. The data base is typically a large, cross-sectional data base of past loans. The methods employed are usually logistic regression, although some of the computer intensive methods such as neural networks can also be used. While historically the problem has not been understood as falling in the forecasting domain, once it is recognized that the effectiveness of the models depends on the periods in which they are built and subsequently applied, the time dimension becomes important.

One of the first studies of models to predict corporate bankruptcy. Though superceded by later work remains an influential reference.

An interesting analysis of forecasters’ behaviour and how they are motivated by other concerns than accuracy.

Uses a data base of analysts’ forecasts of company earnings. More recent forecasts are more accurate, they embody the latest available information. Consensus of the more recent forecasts proved more accurate still (see *Combining*) and therefore serves as a better proxy measure of the market’s expectations of company earnings.

Demonstrates that excess returns that take into account trading costs can be achieved, even when the analysis is restricted to data and models that could have been available to the analyst at the time of producing the forecast, i.e., the methods aim to simulate the situation facing a financial forecaster.

Recent much-cited critical survey of the extensive literature on analyst forecasting.

Remains a current reference point to the burgeoning literature on consumer credit.

Establishing profitably trading in stocks must be done in real-time. Once a forecasting method becomes established, the financial opportunity it presents will disappear. But the authors conclude that innovative new methods should still prove valuable.

Neural nets outperform discriminant analysis, a result that has been replicated in other later applications. Takes a more thorough methodological approach to evaluating the predictions.