Journal of Forecasting, Vol. 15, 495–508 (1996)

Model Uncertainty and Forecast Accuracy

CHRIS CHATFIELD
University of Bath, UK

ABSTRACT

In time-series analysis, a model is rarely pre-specified but rather is typically formulated in an iterative, interactive way using the given time-series data. Unfortunately the properties of the fitted model, and the forecasts from it, are generally calculated *as if the model were known in the first place*. This is theoretically incorrect, as least squares theory, for example, does *not* apply when the same data are used to formulates *and* fit a model. Ignoring prior model selection leads to biases, not only in estimates of model parameters but also in the subsequent construction of prediction intervals. The latter are typically too narrow, partly because they do not allow for model uncertainty. Empirical results also suggest that more complicated models tend to give a better fit but poorer *ex-ante* forecasts. The reasons behind these phenomena are reviewed. When comparing different forecasting models, the BIC is preferred to the AIC for identifying a model on the basis of within-sample fit, but out-of-sample forecasting accuracy provides the real test. Alternative approaches to forecasting, which avoid conditioning on a single model, include Bayesian model averaging and using a forecasting method which is not model-based but which is designed to be adaptable and robust.

KEY WORDS AIC; Bayesian model averaging; BIC; forecasting; model selection; model uncertainty; neural networks; prediction intervals

PRELUDE

Consider the following forecasting problem, which might be regarded as 'typical' of the genre. A statistician is given monthly sales data for the past five years for a particular product and asked to make forecasts for up to 12 months ahead. How would he or she go about this? There is no simple answer in that all decisions depend, for example, on the *context*, on the skill of the analyst, and on the computer software available. The analyst will likely entertain a family of possible models, such as ARIMA or structural state-space models, look at a time plot of the data and at a variety of diagnostic tools such as the autocorrelation function, and then try plausible models within the chosen family. A 'best' model is chosen, and the analyst will then make inferences and forecasts conditional on the selected model being 'true', even though the model has actually been selected from the same data which are now being (re-)used to make predictions. Most (all?) time-series analysts do this sort of thing, but should we?

CCC 0277–6693/96/070495–14

Received June 1995
Revised July 1996

The standard analysis does not take account of the fact that (1) the model has been selected from the same data used to make inferences and predictions and (2) the model may not be correct anyway. Thus the standard analysis ignores the effect of *model uncertainty* which is arguably the most important source of uncertainty. Prediction intervals attempt to allow for the residual variation (though the latter may be underestimated) and may also allow for uncertainty in the estimates of model parameters, but they do not customarily take account of the possibility that the wrong model may have been fitted or that the model may change in the future. This seems unwise and statisticians and forecasters need to address the question as to how model uncertainty will affect forecast accuracy. That is the theme of this paper.

INTRODUCTION

Traditional statistical inference is primarily concerned with the interesting, but rather narrow, problem of estimating, and/or testing hypotheses about, the parameters of a *pre-specified* family of parameter-indexed probability models. Most analysts would agree that their work covers a wider ambit than this, and modern statistical inference is also concerned with *model selection*, *model criticism* and *prediction*. Chatfield (1995a) has argued that statistical inference should be expanded to include the *whole model-building process*. Setting our sights even wider, it should be understood that model building is itself just part of *statistical problem-solving* (e.g. Chatfield, 1995b) where *contextual considerations*, including *objectives*, are critical. Problem solving, like model building, is generally an *iterative* process.

Consider model building as applied to time-series data. *Fitting* a time-series model is usually straightforward nowadays, thanks to a wide range of computer software. Packages also typically carry out a range of routine *diagnostic* checks such as calculating the autocorrelation function of the residuals, and the Box–Ljung 'portmanteau' lack-of-fit statistic. In contrast, *formulating* a sensible time-series model can still be difficult, and yet this aspect of model building has received surprisingly little attention. A time-series model may be specified partly on external subject-matter grounds, or on background theory (e.g. economic theory) or on a model fitted to time series of a similar type. However, these are exceptions rather than the rule, and most time-series models are determined from the given data by an iterative cycle of (1) model formulation, (or model specification), (2) model fitting (or model estimation), (3) model checking (or model validation). This is exemplified by the iterative Box–Jenkins model-building procedure applied to ARIMA models (Box *et al.*, 1994, Section 1.3.2), but nowadays used more generally for most other classes of time-series model. The analyst typically searches over a range of models and selects the model which is 'best' according to some yardstick such as minimizing Akaike's Information Criterion (AIC). Having done this, the analyst proceeds to estimate the parameters of this 'best' model using the *same* techniques as would be used in traditional statistical inference where the model is assumed known *a priori*. Unfortunately this is 'logically unsound and practically misleading' (Zhang, 1992). In particular, least squares theory is known not to apply when the model has, in fact, been selected from the *same* data used for estimation purposes, as happens routinely in time-series analysis, Statisticians have typically ignored this type of problem, partly because it is not clear what else could/should be done. Little theory is available for guidance, and the biases which result when a model is formulated and fitted to the *same* data are not well understood. Such biases are called *model-selection biases* (Chatfield, 1995a).

There are typically three main sources of uncertainty in any problem:

(1) Uncertainty about the structure of the model
(2) Uncertainty about estimates of the model parameters, assuming the model structure is known

(3) Uncertainty about the data even when the model structure and the values of the model parameters are known. This will include unexplained random variation in the observed variables, as well as measurement and recording errors.

The statistical literature has much to say about (2) and (3) but rather little about (1)—see Chatfield (1995a) for a review. Publications relevant to time-series analysts include the collection of papers in Dijkstra (1988), various studies on subset selection in multiple regression (e.g. Miller, 1990; Faraway, 1992; Pötscher and Novak, 1994), Draper's (1995) review of the Bayesian model-averaging approach and the work of Hjorth (1982, 1987, 1989, 1990, 1994) which includes a number of interesting time-series examples. The sparseness of the literature is surprising given that errors arising from *model uncertainty* are likely to be far worse than those arising from other sources. For example, when fitting an autoregressive model, theory tells us about the errors resulting from having estimates of autoregression coefficients rather than their true values, but these errors are likely to be smaller than errors resulting from misspecification, such as omitting a lagged variable by mistake, or failing to include appropriate trend and seasonal terms. Even after the most diligent model-selection process, the analyst cannot be sure that the true model has been found (if there is one—see below) and should bear in mind that a fitted model is, at best, a useful approximation. In view of the seriousness of specification error, it can be argued that it is often inadequate to try to describe uncertainty in the usual way by means of standard errors conditional on the model. Instead we need to find ways of getting more realistic estimates of prediction error, perhaps based on resampling methods, on mixing several models, on empirical experience, or on some sort of *sensitivity analysis*, whereby small changes are made to the model assumptions to see how stable the deductions (including forecasts) from the model are.

This paper discusses various aspects of model uncertainty in regard to time-series analysis and forecasting. Data-dependent model specification searches can lead to non-trivial biases, both in estimates of model parameters and in the ensuing forecasts. Methods for tackling the problem are discussed but unfortunately there is no simple general solution. The main message of this paper is that, when a time-series model is formulated and fitted to the same data, inferences and forecasts made from it will be biased and (seriously) over-optimistic when they ignore the prior model-selection process.

EXAMPLES

Given the difficulties of proving general theoretical results about the effects of model selection on subsequent inferences, the use of specific examples, perhaps employing simulation, can be particularly enlightening to demonstrate undesirable effects. Examples on regression modelling are given by Miller (1990) and Chatfield (1995a, Examples 2.2 and 2.3), while examples particularly relevant to time-series-analysis and forecasting are given by Hjorth (1987, Examples 5 and 7, 1994, Example 2.2) and Chatfield (1995a, Examples 2.4 and 2.5). This section illustrates the difficulties with theory with some additional comments on Chatfield's (1995a) Example 2.4, and then demonstrates empirical difficulties in practice with an example fitting neural network models.

Example 1. Fitting an AR(1) model. Consider the first-order autoregressive (AR(1)) time-series model, namely:

$$X_t = \alpha X_{t-1} + \varepsilon_t$$

where $|\alpha| < 1$ for stationary and $\{\varepsilon_t\}$ are i.i.d. $N(0, \sigma^2)$. Given a sample of data, it is straightforward to fit an AR(1) model and estimate α. However, in practice with real data, the analyst is

unlikely to know *a priori* that the AR(1) model really is appropriate. A simple (perhaps oversimplified) identification procedure consists of calculating the first-order autocorrelation coefficient, r_1, and fitting the AR(1) model if, and only if, r_1 is significantly different from zero. What this procedure does is to eliminate the possibility of getting 'small' values of $\hat{\alpha}$ which correspond to 'small' values of r_1. Thus the resulting (conditional) estimate of α turns out to be biased. Large values of $\hat{\alpha}$ are more likely to occur as can readily be demonstrated theoretically or by simulation.

This example emphasizes that, in assessing bias, the analyst must be clear exactly what the inference is conditioned on. Theory tells us about the unconditional expectation of $\hat{\alpha}$ where an AR(1) model is always fitted. However, if the model selection is taken into account, then the appropriate expectation is $E(\hat{\alpha} \mid r_1$ is significant). There is also another estimator which is arguably of interest, namely

$$\hat{\alpha}_{PT} = \begin{cases} \hat{\alpha} & r_1 \text{ is significant} \\ 0 & \text{otherwise} \end{cases}$$

This estimator can be recognized as a simple example of what econometricians call a *pre-test* estimator (e.g. Judge and Bock, 1978). It arises by recognizing that, when r_1 is not significant and an AR(1) model is not fitted, this could be regarded as fitting an AR(1) model with $\alpha = 0$. It is immediately apparent that the three quantities $E(\hat{\alpha})$, $E(\hat{\alpha} \mid r_1$ is significant) and $E(\hat{\alpha}_{PT})$ will generally not be equal. Moreover, it is clear that the three estimators will have different sampling distributions and hence different variances. Given that estimators of model parameters are biased, it is not surprising to find that the estimated residual standard deviation is also likely to be biased (see further comments on this point in the next section).

Of course, the above model-selection procedure is simpler than would normally be the case in time-series analysis. More typically the analyst will inspect autocorrelations and partial autocorrelations of a suitably differenced series, allow the removal or adjustment of outliers and entertain all ARIMA models up to say third order. Choosing a 'best' model from such a wide set of possibilities seems likely to make model selection biases even larger.

While this example has focused on estimates of model parameters, the results are, of course, relevant to forecasting since prediction intervals are calculated conditional on the fitted model. If estimates of model parameters (including the 'error' variance—see below) are biased, then the resulting prediction intervals can also be expected to be biased, and this is indeed the case. Unfortunately there is no easy general way to quantify these biases and progress seems likely to be made primarily by simulation and by empirical experience.

Example 2. Empirical results when fitting neural network models. Neural network (NN) models have recently been applied by several authors to time-series analysis and forecasting problems (e.g. de Groot and Würtz, 1991; Hill *et al.*, 1994). NNs can be thought of as a type of non-linear regression model and an introductory account is given by Ripley (1993). This class of models allows the analyst to fit a large number of parameters and try many different architectures which means that a good (within-sample) fit can usually be obtained. However, there is a real danger of *overfitting*, and their forecasting ability is still unproven (Chatfield, 1993b) despite some media 'hype'. Indeed White (1994) reported a large study of economic series where the random walk model often outperformed neural nets in out-of-sample forecasts. Faraway and Chatfield (1995) applied a variety of NN models to the famous airline data (Box *et al.*, 1994, Series G) and the results relating to out-of-sample forecast accuracy are further developed here. All NN models were of the usual feedforward type with one hidden layer of neurons. The input variables were

Table I. Comparison of the fit and predictions for various NN models for the airline data using the first 132 observations as the training set

Lags	No. of hidden neurons	No. of pars.	Fit $\hat{\sigma}$	Fit AIC	Fit BIC	Predictions SS_{MS}	Predictions SS_{1S}	Predictions $\hat{\sigma}_{pred}$
1, 12, 13	1	6	0.102	−537.1	−514.4	0.334	0.504	0.20
1, 12, 13	2	11	0.098	−543.1	−501.4	0.329	0.503	0.20
1, 12, 13	4	21	0.093	−546.8	−467.4	0.538	0.621	0.23
1, 12	2	9	0.144	−456.3	−422.2	0.351	0.344	0.17
1, 12	4	17	0.145	−447.7	−383.5	0.376	0.443	0.19
1, 12	10	41	0.150	−423.7	−268.4	0.508	0.592	0.22
1, 2, 12	2	11	0.141	−459.4	−417.7	0.339	0.291	0.16
1, 2, 12	4	21	0.139	−454.6	−375.1	6.820	1.032	0.29
1, 2, 12, 13	2	13	0.097	−543.5	−494.4	0.374	0.519	0.21
1, 2, 12, 13	4	25	0.093	−543.1	−448.7	0.339	0.517	0.21
1–13	2	31	0.091	−544.8	−427.6	1.078	0.709	0.24
1–13	4	61	0.067	−605.1	−374.6	4.116	1.122	0.31

the values of the given variable (the number of airline passengers) at selected lags so that attention was restricted to univariate forecasts where forecasts of X_t depend only on past values of the series. In order to avoid 'silly' models, the values at lags one and twelve were always included. The logistic activation function was used at the hidden layer and the identity activation function at the output stage. Initially the models were fitted to the first eleven years of data (the training set in NN jargon) and the last year's data (the test set) was used for making genuine out-of-sample forecast comparisons.

Selected results are shown in Table I, where $\hat{\sigma}$ = estimated residual standard deviation for the training set, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are defined in the usual way, and SS_{MS}, SS_{1S} are the sum of squares of multi-step and one-step-ahead (out-of-sample) forecasts made of the last year's data. The multi-step forecasts were all made in month 132. The prediction error standard deviation, denoted by $\hat{\sigma}_{pred}$, is estimated by $\sqrt{(SS_{1S}/12)}$.

As the number of lagged variables and the number of hidden neurons are increased, the number of parameters in the fitted NN model increases alarmingly. Several models have in excess of 20 parameters even though the number of observations in the training set is only 132. Many (most?) analysts would guess that it would be unwise to have more than about 10 parameters with so few observations. Inspection of Table I bears out this view when the accuracy of the predictions is assessed.

Generally, Table I demonstrates that the more parameters are fitted, the lower will be the value of $\hat{\sigma}$, as would be expected. The relationship with AIC, which penalizes the addition of extra parameters, is less clear-cut. However, the minimum value of AIC for the models tabulated is for a 61-parameter model. The model giving the next lowest AIC is a 21-parameter model. In contrast, the model giving the lowest BIC, which penalizes extra parameters more severely than the AIC, is for a 6-parameter model. Thus the use of AIC or BIC leads to completely different model choices.

Turning from fit to predictions, Table I tells us that getting a good fit—meaning a low value of $\hat{\sigma}$—is a poor guide to getting good predictions. Indeed models with *smaller* numbers of parameters generally give *better* (out-of-sample) predictions even though they may give a worse fit than less parsimonious models. In particular the model selected as 'best' by BIC gives much

better predictions than the model selected as 'best' by AIC. This finding suggests that *BIC is a better criterion than AIC* for choosing a model based on within-sample fit in order to get good out-of-sample predictions.

Table I also allows us to compare fit with forecast accuracy more generally. The results might be described as alarming. The within-sample estimate of the error standard deviation (i.e. $\hat{\sigma}$) is typically much less than the (out-of-sample) one-step-ahead prediction error standard deviation (i.e. $\hat{\sigma}_{\text{pred}}$). For the better models (with a small number of parameters and low BIC), we find $\hat{\sigma} \simeq 0.1$ and $\hat{\sigma}_{\text{pred}} \simeq 0.2$. So the latter is about double the former. For the more dubious models (with low $\hat{\sigma}$) and low AIC, but with higher numbers of parameters), the ratio of $\hat{\sigma}_{\text{pred}}$ to $\hat{\sigma}$ becomes disturbingly large—for example, it rises to 4.6 for the 61-parameter model with the 'best' AIC.

The above results report what happens when different NN models are fitted to the airline data using the first 132 observations as the training set. Qualitatively similar results were found using the first 126 observations as the training set and also using a completely different data-set, namely the Chatfield–Prothero sales data (Chatfield and Faraway, 1996). Moreover when fitting Box–Jenkins seasonal ARIMA models to the data, it was also found that $\hat{\sigma}_{\text{pred}}$ was typically, at best, twice as large as $\hat{\sigma}$.

Why is $\hat{\sigma}_{\text{pred}}$ so much larger than $\hat{\sigma}$ in these cases, and is this a finding which generalizes to other data sets and models? Empirical experience (e.g. Meese and Geweke, 1984; Chatfield, 1993a, Section 6; Fildes and Makridakis, 1995, p. 295) suggests that the answer to the second question is 'Yes'. *Out-of-sample forecast accuracy is generally (much) worse than would be expected from within-sample fit.* Some theoretical results, such as the optimism principle (see the next section), help to explain the above. While there are other contributory factors, it seems likely that model uncertainty is the prime cause. Either the wrong model is identified or the true model is changing through time in a way which is not captured by the forecasting mechanism. Perhaps the most important consequence of the above is that *comparisons of different forecasting models and methods should preferably be made on the basis of out-of-sample predictions.*

MODEL BUILDING IN TIME-SERIES ANALYSIS

Constructing a plausible model is an important ingredient of time-series analysis, and hence of forecasting. Many years ago it may have been true that a *single* model was typically fitted to a given set of data. Nowadays the increase in computing power has completely changed the way in which time-series analysis is typically carried out (not necessarily for the better!). A model is often selected from a wide class of models by optimizing a statistic such as adjusted-R^2, AIC or BIC, and there is a large literature on model selection in time-series analysis—see, for example, Gooijer *et al.* (1985) and Choi (1992). As well as choosing from a wide class of models, the data-analysis procedure may also involve strategies such as (1) excluding, downweighting or otherwise adjusting outliers; (2) transforming one or more variables, for example to achieve normality and/or constant residual variance. For example, a time-series analyst might start an analysis by entertaining the class of ARIMA(p, d, q) models for say $0 \leqslant p, d, q \leqslant 2$. This looks fairly innocent, but actually allows a total of 27 possible models even before considering outliers or transformations, quite apart from additional possibilities such as finding seasonality (suggesting a seasonal ARIMA model) or non-linearities (suggesting a completely different class of models). Clearly the analyst may in effect consider tens or even hundreds of models.

In econometrics, the situation is further complicated by the common practice of pretesting various hypotheses, such as tests for a unit root, for autocorrelated residuals, or for the presence

of a change point. The desire to carry out such tests indicates awareness of model uncertainty, but is viewed with suspicion by some statisticians, especially when a large number of tests is performed. Why for example should the presence of a unit root be taken as the null hypothesis? We do not pursue this topic here (see Chatfield, 1995a, Example 2.5; Ormerod, 1996, especially Section 4), except to note that inference following model-testing is biased, and that testing implicitly assumes the existence of a 'true' model (see below).

Given the wide choice of models, the chance of choosing the correct one is slim, even assuming there really is a 'true' model which is in the set of models entertained (called the *M-closed view* by Bernado and Smith (1994, Section 6.1.2)). Most model builders would admit (privately at least!) that they do not really believe there is a true model (see Chatfield, 1995a, Section 3.1 and the discussion which follows that paper). Rather, a model is seen as a useful description of the given data which provides an adequate approximation for the task at hand. Here the *context* and the *objectives* are key factors in making such a judgement.

There are various alternatives to the rigidity of assuming there is a 'true' model. There is, for example, increasing interest in *local* models with changing parameters (such as structural or state-space models) rather than *global* models with constant parameters (such as ARIMA models). As one example, the use of regression on time to model deterministic trends has fallen out of favour as compared with techniques which fit a local trend. Local models are often fitted by some sort of updating procedure, such as the *Kalman filter*, which are easy to apply using a computer, and the use of such techniques seems likely to increase. In this regard it is interesting to note that simple exponential smoothing (which is a very simple type of Kalman filter) is optimal for two models which appear to be of a completely different type, namely the ARIMA(0, 1, 1) model, which has constant parameters, and the steady (state-space) model (e.g. Chatfield, 1996, Section 10.1.1), which allows the local mean level to drift through time. Given that a 'true' model probably does not exist, there is much to be said for choosing a forecasting method, not because it is optimal for a particular model but rather because it can adapt to changes and works well in practice. Thus empirical results suggest that the seasonal version of exponential smoothing, called Holt–Winters, is robust to model changes (Chen, 1994).

A second alternative to assuming the existence of a single 'true' model is to allow the possibility that there may be *more than one* model which may be regarded as a sufficiently close approximation to the given data for the required objective (e.g. Poskitt and Tremayne, 1987). The notion of having more than one model is a key element of the Bayesian model-averaging approach and is implicit in the *combination* of forecasts—see the next section.

A third alternative is to use different models to describe different parts of the data, as may seem appropriate, for example, when the properties of recent observations differ markedly from those of earlier values.

Finally, a fourth alternative is to use different models for different lead times. It has been established empirically (e.g. Meese and Geweke, 1984; Gersch and Kitagawa, 1983) that the model which works best for, say, short-term forecasting may not be so good for longer lead times. The criterion for choosing a model needs to be matched to the given application rather than relying on theoretical results which assume the true model is known. In particular, when k-steps-ahead forecasts are required, it may be advisable to fit a model by minimizing prediction errors k-steps-ahead, rather than one step ahead (e.g. Stoica and Nehorai, 1989; Tiao and Xu, 1993).

If we ignore the above alternatives and behave as if we believe the best-fitting model to be true, then problems inevitably arise. It is indeed illogical to admit model uncertainty by searching for a 'best' model, but then ignore this uncertainty by making inferences and predictions as if certain that the best-fitting model is actually true. Statistical theory has not kept pace with the

computer-led revolution in statistical practice and there has been little progress in understanding inference after model-selection, even though widespread biases arise.

Chatfield (1995a) reviews the rather limited research on inference after model selection, both in regard to (1) assessing the size of model-selection biases; and (2) methods of overcoming or circumventing the problem. One key message is that *the properties of an estimator may depend not only on the selected model but also on the selection process.* The use of a model-selection statistic essentially partitions the sample space into disjoint subsets, each of which leads to a different model. This vantage point enables the derivation of various inequalities regarding the expectation of the optimized statistic and provides a theoretical justification for what Picard and Cook (1984) call 'the Optimism Principle', namely that the fitting of a model typically gives optimistic results in that performance on new data is, on average, worse than on the original data. As Hjorth (1989) says, 'it is perhaps not surprising that selection minimizing a criterion will cause underestimation of this criterion'. In particular, if a time-series model is selected by minimizing mean square prediction error (MSPE), then the Optimism Principle explains why the within-sample fit of a best-fitting time-series model is typically better then out-of-sample forecasts. This is reminiscent of the *shrinkage* effect in regression (e.g. Copas, 1983), and of experience with discriminant analysis where discrimination on a new set of data (a *test* sample) is typically worse than for the data used to construct the discrimination rule (the *training* sample).

The difficulty in making theoretical progress has led to a number of simulation studies and the use of a variety of computational procedures such as *resampling*, *bootstrapping*, *jackknifing*, *cross-validation*, and *data-splitting*. The results were reviewed by Chatfield (1995, Section 4) and only brief notes will be given here as they relate more to parameter estimation. As one example, Miller (1990, p. 160) found alarmingly large biases, of the order of one to two standard errors, in the estimates of regression coefficients when using subset selection methods in multiple regression. Hjorth (1987, Example 5) simulated data from an ARMA(1, 1) model, but found that the correct ARMA model was identified in only 28 out of 500 series. The properties of the ARMA(1, 1) parameter estimates for the 28 series differed greatly from those arising when the model was fitted to all 500 series. Furthermore, the average estimated MSPE was *less than one-third* of the true MSPE for the model which was actually fitted. Pötscher and Novak (1994) simulated various MA and AR models but selected the order from the data. They found the 'the distribution of post-model-selection estimators frequently differs drastically from the distribution of LS estimates based on a model of fixed order'. It is sad that results such as these are largely ignored in practice. Computational methods can be used to study the effects of data-dependent model-selection provided the model-selection procedure is clearly defined (which it will not always be in practice). For example, Faraway (1992) simulated the actions taken during regression analysis, including the handling of outliers and transformations. In time-series analysis, resampling is particularly tricky because of the ordered nature of the data and because one has to avoid conditioning on the fitted model (which would not reflect model uncertainty). Nevertheless, careful bootstrapping can overcome much of the bias due to model uncertainty.

A computational technique which is used much more widely in time-series analysis and forecasting is *data splitting*. This involves splitting the series into two parts, fitting the model to the first part (sometimes called the *construction* or *training* sample) and using the second part (sometimes called the *hold-out*, *test* or *validation* sample) to check inferences and predictions. One problem is deciding *how* to split the data (e.g. see Picard and Cook, 1984), but there are no general guidelines on this. Moreover, fitting a model to just part of a time series will result in a loss of efficiency and so some compensatory effect is needed. Unfortunately Faraway (1992) shows that, in regression modelling, data splitting may increase the variability in estimates without the reward of eliminating bias. This result may well generalize to time-series modelling.

For statistical applications other than time-series analysis, hold-out samples are a poor substitute for taking a true replicate sample (Chatfield, 1995a, Section 6). However, it is usually impossible to replicate time-series data (except by waiting for many time periods which is hardly practicable). Thus, despite its drawbacks, data splitting can be used in forecasting to provide a salutary check on over-optimistic forecasts. However, note that if the hold-out sample is used to help select the 'best' model (e.g. by picking the model which gives the best forecasts of the holdout sample, rather than the best fit to the construction sample), then it is no longer a genuine hold-out sample and will no longer provide an independent check.

We have concentrated on inferential biases, but it should be noted that the literature on *model checking* is also questionable. It is theoretically desirable for a hypothesis to be validated on a second confirmatory sample but this is usually impossible in time-series analysis. Rather, diagnostic checks are carried out on the *same data* used to fit the model. Now diagnostic tests assume the model is specified *a priori* and calculate a *P*-value as Probability(more extreme result than the one obtained | model is true). But if the model is formulated, fitted and checked using the same data, then we should really calculate Probability(more extreme result than the one obtained | model has been selected as 'best' by the model-selection procedure). It is not clear in general how this can be calculated. However, it is clear that the good fit of a 'best-fitting' model should not be surprising, and empirical experience tells us that diagnostic checks hardly ever reject the best-fitting time-series model precisely because it is the best fit!

FORECASTING AND MODEL UNCERTAINTY

This section looks more directly at the effect of model uncertainty on the choice of forecasting method. Much of statistical inference is concerned with estimating *unobservable* quantities, such as population parameters, where the analyst may never know if the inferences are 'good' since the estimates cannot be compared directly with the truth. However, time-series analysis involves the prediction of *observable* quantities, which provides an excellent opportunity to check or *calibrate* a model (Geisser, 1993). It is therefore sad that empirical findings are too often ignored by theoreticians who continue to derive results on inference and forecasting which assume the existence of a true, known model (Fildes and Makridakis, 1995).

Model uncertainty is clearly crucial in forecasting since, if the analyst uses an inappropriate model, then forecasts will be less accurate. It is tempting to think that one can simply fit more and more terms to get an adequate approximation but that does not work. A more complicated model may reduce bias (though not if unnecessary terms are included), but may also increase variance, because more parameters have to be estimated (Breiman, 1992, p. 738). For example, Davies and Newbold (1980) show that, although an MA(1) model can be approximated arbitrarily closely by an AR model of high order, the effect of having to estimate additional parameters from finite samples is that forecast error variance gets worse for higher-order models. More generally, inexperienced analysts may intuitively expect more complicated models to give better forecasts. However, empirical evidence suggests the reverse and many examples could be given. Example 2 above is one such, while a second recent example (Collopy *et al.*, 1994) shows that simple extrapolation outperforms a more complicated diffusion model when forecasting spending on information systems. The Optimism Principle introduced in the previous section provides one explanation as to why a more complicated model may give a better fit but worse predictions. This distinction between within-sample fit and out-of-sample forecasts reminds us to ensure that all forecasting comparisons of different models and methods are made on genuine *ex-ante* (or out-of-sample) forecasts—see Armstrong (1985, p.p. 338–9).

The principle of *parsimony* says that the smallest possible number of parameters should be used so as to give an adequate representation of the data and this principle should be borne in mind when model specification and selection is taking place. The more complicated the model, the more possibilities there will be for departures from model assumptions. The dangers of overfitting, so important when constructing multiple regression and autoregressive models, are well illustrated by the neural network models in Example 2. Unfortunately, these dangers are not always heeded.

In assessing forecasts, it is also important to realize that models which are mathematically very different, and which give very different long-term predictions, may be virtually indistinguishable in terms of their fit to a set of data. For example, for some near-non-stationary sets of data, an AR(1) model with a parameter close to unity will give a similar fit and similar one-step-ahead forecasts to those from a (non-stationary) random walk model. However, the forecasts many steps ahead from these two models are quite different. The limiting point forecasts are respectively equal to the mean for the AR(1) model and to the latest observation for the random walk, while the limiting prediction error variance is finite for the stationary model but infinite for the non-stationary model. Getting the 'wrong' form of differencing makes little difference to short-term point forecasts, but, for long-term forecasts, the fiction that there is no model uncertainty is far from innocuous. Similar remarks apply to extrapolating from any model. An instructive example concerns the Challenger space shuttle disaster data where it is hard to distinguish between several models in terms of fit, but where the long-term extrapolations are very different (Draper, 1995, Section 6.2). Forecasting the spread of AIDS provides another example (Draper, 1995, reply to the discussion).

A consequence of formulating and fitting a model to the same data is that MSPE is underestimated. Partly as a result, *prediction intervals will generally be too narrow*. Empirical studies have shown that nominal 95% prediction intervals will typically contain (much) less than 95% of actual future observations for a variety of reasons (see Chatfield, 1993a, Section 6). Model uncertainty is one important reason, not only for the way it leads to underestimates of MSPE but also because the model may be incorrectly identified or may change through time. There is an alarming tendency for analysts to think that narrow intervals are 'good' when wider ones may well reflect model uncertainty better. Draper (1995) presents an instructive example concerning forecasts of the price of oil. Ten models were entertained which gave a wide range of point forecasts that were nevertheless all well away from the actual values which resulted. There were also large differences in the prediction error variances. A model uncertainty audit suggested that only about 20% of the overall predictive variance could be attributed to uncertainty about the future of oil prices conditional on the selected model and on the assumptions (the scenario) made about the future. Yet the latter portion is all that would normally be taken into consideration. Other case studies (e.g. Wallis, 1986) which have examined the decomposition of forecast errors have also found that the contribution due to model specification uncertainty can be substantial.

Draper (1995) went on to consider the oil price example from the point of view of *Bayesian model averaging*. This technique should appeal not only to Bayesians but also to any 'broad-minded' statistician. The key to its success lies in not having to choose a single 'best' model but rather in averaging over several plausible competing models which are entertained with appropriate prior probabilities. The data are then used to evaluate posterior probabilities for the different models. Models with 'low' posterior probabilities may be discarded to keep the problem manageable, and then a weighted sum of the predictions from the remaining competing models is calculated. Under certain assumptions, the combined forecast from Bayesian model averaging will have a lower MSPE in the long run that the forecasts from any of the individual models.

The idea behind Bayesian model averaging suggests that the way time-series models are customarily fitted should be reconsidered. For example, suppose it is desired to fit an AR model to a set of data. The appropriate order is usually unknown *a priori*, and the approach generally adopted at present is to assess the order by minimizing a criterion such as AIC, and then estimate model parameters and make predictions conditional on the selected order being correct. The alternative suggested by Bayesian model averaging is to recognize that it is unlikely that an AR model is the true model, and, even if it is, the correct order may not be chosen. This suggests approximating the data by a mixture of AR models of different orders, rather than relying on a single model of fixed order. Prior probabilities for different orders would need to be assessed and posterior probabilities evaluated from the data, for example by Gibbs sampling (Barnett *et al.*, 1993). Successful applications of Bayesian model-averaging to AR processes are reported by Schervish and Tsay (1988) and Le *et al.* (1993).

Despite its promise, there are difficulties in applying Bayesian model averaging. The calculation of posterior probabilities from the prior probabilities requires the computation of Bayes factors which may not be easy, even in this computer age. Moreover, prior probabilities for the different models have to be specified and this is not easy, especially when some models are entertained only *after* looking at the data. Finally Bayesian model averaging does not lead to a simple model, and although this doesn't matter for forecasting purposes, it does matter for the purposes of description and interpretation.

The general idea of mixing several models, rather than having to use a single 'best' model, is attractive and is the idea behind the use of multi-process or mixture models in Bayesian forecasting (West and Harrison, 1989, Chapter 12). Two other forecasting approaches, relevant to our discussion, also come to mind here. In long-range forecasting, *scenario analysis* (e.g. Schoemaker, 1991) is often used. Here a variety of different assumptions are made about the future giving a range of forecasts, rather than just one. Each forecast is linked clearly to the assumptions it depends on, and the spread of forecasts should clarify the extent of model uncertainty. The aim is to allow organizations to make contingency plans for different possible futures. A completely different type of strategy arises from *combining forecasts* (in a non-Bayesian way). Suppose you have produced forecasts by several different methods. Then it has been established empirically that a weighted linear combination of these forecasts will often be more accurate on average then any of the individual forecasts (e.g. Clemen, 1989). A simple average is often as good as anything. Unfortunately, the client does not get a simple model to describe the data, and the stochastic properties of the combined forecast may also be unclear.

Instead of mixing several models, the final possibility mentioned here is to use a forecasting *method* which is not model-based but which is designed to readily adapt to changes in the underlying model. The various forms of exponential smoothing come into this category. Although optimal for particular models, their main justification lies in giving good empirical results for a variety of data sets.

SUMMARY AND DISCUSSION

The theory of inference regarding parameter estimation generally assumes that the true model for a given set of data is known and pre-specified. In practice a time-series model is usually formulated from the data, and many models may be entertained. A single model is usually selected as the 'winner' even when other models give nearly as good a fit. Given that the wrong model may be selected or that a 'true' model may not exist anyway, it follows that *model*

uncertainty is present in most real problems. It is therefore surprising that the topic has received so little attention from forecasters.

When inference and prediction follow data-dependent model selection, the following general points can be made:

(1) Least-squares theory does not apply when the same data are used to formulate and fit a model. Yet time-series textbooks (including my own (Chatfield, 1996)!) habitually ignore this point.
(2) After model selection, estimates of model parameters and of the residual variance are likely to be biased.
(3) Models with more parameters generally give a better fit, but may give worse out-of-sample predictions. In comparing the fit of different models, the BIC is preferred to the AIC so as to penalize adequately the introduction of additional parameters.
(4) The analyst typically thinks the fit is better than it really is (the Optimism Principle), and prediction intervals are generally too narrow, partly because residual variance tends to be underestimated and partly because they fail to take full account of model uncertainty. Moreover diagnostic checks rarely reject the best-fitting model precisely because it *is* the best fit!
(5) The frequentist approach does not adapt naturally to cope with model uncertainty, though some progress can be made with resampling and other computational methods. Bayesian model averaging offers a promising alternative approach even to analysts who are not Bayesian. However, difficulties arise whichever approach is adopted, and there appears to be no simple general theoretical 'fix'.

So how should the results in this paper affect the way a forecaster proceeds? Despite the difficulties in making general recommendations, the following advice will be given in addition to the well-tried counsel to clarify the objectives of the forecasting exercise, find out exactly how a forecast will actually be used, ascertain whether a model is required for descriptive purposes as well as for prediction, and ask questions to get background information as to a suitable class of models.

(i) Be alert to the insidious presence of model uncertainty but be aware that there is no simple way of overcoming the problem.
(ii) Realize that the computer-based revolution in time-series model-fitting means that the analyst typically looks at a (very) large number of models. This leads to biases when inference follows data-based model-selection.
(iii) Realize that more complicated models, while generally giving a better fit, do not necessarily give better out-of-sample forecasts. Use the BIC, rather than the AIC, to select a model on the basis of within-sample fit. Realize that forecasts from a best-fitting model will generally not be as good as expected, that prediction intervals will generally be too narrow and that the real test of a forecasting model or method is its out-of-sample forecasting ability.
(iv) Clonsider the following alternatives to the use of a single best model; (a) use a mixture of models; (b) use a forecasting method which is not model-based but which is adaptive and robust.

ACKNOWLEDGEMENTS

The results in Example 2 were developed jointly with Dr Julian Faraway, Department of Statistics, University of Michigan, USA. I thank the referees for some constructive suggestions.

This paper is based on an invited talk given to an ESRC workshop on 'Model Complexity and Forecast Accuracy' at the Lancaster Centre for Forecasting, April 1995.

REFERENCES

Armstrong, J. S., *Long-Range Forecasting*, 2nd edn, New York: Wiley, 1985.

Barnett, G., Kohn, R. and Sheather, S., 'Bayesian estimation of an autoregressive model using Markov chain Monte Carlo', Working paper 93-012, Australian Graduate School of Management, 1993.

Bernado, J. M. and Smith, A. F. M. *Bayesian Theory*, Chichester: Wiley, 1994.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., *Time Series Analysis, Forecasting and Control*, 3rd edn. Englewood Cliffs, NJ: Prentice Hall, 1994.

Breiman, L., 'The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error', *J. Am. Statist. Ass.*, **87** (1992), 738–754.

Chatfield, C., 'Calculating interval forecasts (with discussion)', *J. Business and Economic Statist.*, **11** (1993a), 121–144.

Chatfield, C., 'Neural networks: Forecasting breakthrough or passing fad?' *Int. J. Forecasting*, **9** (1993b), 1–3.

Chatfield, C., 'Model uncertainty, data mining and statistical inference (with discussion)', *J. R. Statist. Soc. A*, **158** (1995a), 419–466.

Chatfield, C., *Problem Solving: A Statistician's Guide*, 2nd edn, London: Chapman and Hall, 1995b.

Chatfield, C., *The Analysis of Time Series*, 5th edn, London: Chapman & Hall, 1996.

Chatfield, C. and Faraway, J., 'Forecasting sales data with neural nets: A case study', *Recherche et Applications en Marketing*, **11** (1996), 29–41.

Chen, C., 'Some statistical properties of the Holt–Winters seasonal forecasting method', Paper presented to the 14th Int. Symposium on Forecasting, Stockholm, 1994.

Choi, B., *ARMA Model Identification*, New York: Springer-Verlag, 1992.

Clemen, R. T., 'Combining forecasts: A review and annotated bibliography', *Int. J. Forecasting*, **5** (1989), 559–583.

Collopy, F., Adya, M. and Armstrong, J. S., 'Principles for examining predictive validity: The case of information systems spending forecasts', *Information Systems Research*, **5** (1994), 170–179.

Copas, J. B., 'Regression, prediction and shrinkage (with discussion)', *J. R. Statist. Soc., B*, **45** (1983), 311–354.

Davies, N. and Newbold, P., 'Forecasting with misspecified models', *Applied Statistics*, **29** (1980), 87–92.

de Groot, C. and Würtz, D., 'Analysis of univariate time series with connectionist nets: A case study of two classical examples', *Neurocomputing*, **3** (1991), 177–192.

Dijkstra, T. K. (ed.), *On Model Uncertainty and its Statistical Implications*, Berlin: Springer-Verlag, 1988.

Draper, D., 'Assessment and propagation of model uncertainty', *J. R. Statist. Soc. B*, **57** (1995) 45–97.

Faraway, J. J., 'On the cost of data analysis', *J. Computational and Graphical Statistics*, **1** (1992), 213–229.

Faraway, J. J. and Chatfield, C., 'Time series forecasting with neural networks: A case study', Statistics Group Research Report 95:06, University of Bath, 1995.

Fildes, R. and Makridakis, S., 'The impact of empirical accuracy studies on time series analysis and forecasting', *Int. Statist. Rev.*, **63** (1995), 289–308.

Geisser, S., *Predictive Inference: An Introduction*, New York: Chapman and Hall, 1993.

Gersch, W. and Kitagawa, G., 'The prediction of time series with trends and seasonalities', *J. Business and Economic Statist.*, **1** (1983), 253–264.

Gooijer, J. G. de, Abraham, B., Gould, A. and Robinson, L., 'Methods for determining the order of an autoregressive-moving average process: A survey', *Int. Statist. Rev.*, **53** (1985), 301–329.

Hill, T., Marquez, L., O'Connor, M. and Remus, W., 'Artificial neural network models for forecasting and decision making', *Int. J. of Forecasting*, **10** (1994), 5–15.

Hjorth, U., 'Model selection and forward validation', *Scandinavian J. Statist.*, **9** (1982), 95–105.

Hjorth, U., 'On model selection in the computer age', Technical report no. LiTH-MAT-R-87-08, Linköping University, Sweden, 1987.

Hjorth, U., 'On model selection in the computer age', *J. Statistical Planning and Inference*, **23** (1989), 101–115.

Hjorth, U., 'Model selection needs resampling methods', Technical report no. LiTH-MAT-R-1990-12, Linköping University, Sweden, 1990.

Hjorth, U., *Computer Intensive Statistical Methods—Validation Model Selection and Bootstrap*, London: Chapman and Hall, 1994.

Judge, G. G. and Bock, M. E., *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*, Amsterdam: North-Holland, 1978.

Le, N. D., Raftery, A. E. and Martin, R. D., 'Robust model comparison for autoregressive processes with robust Bayes factors', Technical report no. 123. Department of Statistics, University of British Columbia, 1993.

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

Miller, A. J., *Subset Selection in Regression*, London: Chapman and Hall, 1990.

Ormerod, P., Paper prepared for the *Keynes, Knowledge and Uncertainty* project second international conference, University of Leeds, UK, March 1996.

Picard, R. R. and Cook, R. D., 'Cross-validation of regression models', *J. Am. Statist. Ass.*, **79** (1984), 575–583.

Poskitt, D. S. and Tremayne, A. R., 'Determining a portfolio of linear time series models', *Biometrika*, **74** (1987), 125–137.

Pötscher, B. M. and Novak, A. J., 'The distribution of estimators after model selection: Large and small sample results', Department of Statistics Working Paper, University of Vienna, 1994.

Ripley, B. D., 'Statistical aspects of neural networks', in Barndorff-Nielsen, O., Jensen, J. and Kendall, W. (eds), *Chaos and Neural Networks—Statistical and Probabilistic Aspects*, pp. 40–123, London: Chapman and Hall, 1993.

Schervish, M. J. and Tsay, R. S., 'Bayesian modeling and forecasting in autoregressive models', in Spall, J. C. (ed.), *Bayesian Analysis of Time Series and Dynamic Models*, pp. 23–52, New York: Marcel Dekker, 1988.

Schoemaker, P. J. H., 'When and how to use scenario planning: A heuristic approach with illustrations', *J. Forecasting*, **10** (1991), 549–564.

Stoica, P. and Nehorai, A., 'On multistep prediction error methods for time series models', *J. of Forecasting*, **8** (1989), 357–368.

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

Wallis, K. F. (ed.), *Models of the U.K. Economy*, Oxford: Oxford University Press, 1986.

West, M. and Harrison, P. J., *Bayesian Forecasting and Dynamic Linear Models*, New York: Springer-Verlag, 1989.

White, H., 'Can neural networks forecast in the big league? Comparing forecasts to the pros', Key Note Address to the 14th Int. Symposium on Forecasting, Stockholm, 1994.

Zhang, P., 'Inference after variable selection in linear regression models', *Biometrika*, **79** (1992), 741–746.

Author's biography:

Chris Chatfield is Reader in Statistics, School of Mathematical Sciences, University of Bath. His research interests include all aspects of time-series analysis, especially forecasting and the use of neural networks, as well as the more general strategical issues involved in model formulation and problem-solving.

Author's address:

Chris Chatfield, School of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.