Prediction Intervals

Chris Chatfield

Department of Mathematical Sciences,

University of Bath

(Final version: May 1998)

ABSTRACT

Computing prediction intervals (P.I.s) is an important part of the forecasting process intended to indicate the likely uncertainty in point forecasts. The commonest method of calculating P.I.s is to use theoretical formulae conditional on a best-fitting model. If a normality assumption is used, it needs to be checked. Alternative computational procedures that are not so dependent on a fitted model include the use of empirically based and resampling methods. Some so-called approximate formulae should be avoided. P.I.s tend to be too narrow because out-of-sample forecast accuracy is often poorer than would be expected from within-sample fit, particularly for P.I.s calculated conditional on a model fitted to past data. Reasons for this include uncertainty about the model and a changing environment. Ways of overcoming these problems include using a mixture of models with a Bayesian approach and using a forecasting method that is designed to be robust to changes in the underlying model.

Keywords: Bayesian forecasting; Bootstrapping; Box-Jenkins method; Holt-Winters method; Prediction intervals; Resampling

INTRODUCTION

Predictions are often expressed as single numbers, called *point forecasts*, which give no guidance as to their likely accuracy. They may even be given with an unreasonably high number of significant digits implying spurious accuracy! Now point forecasts sometimes appear adequate, as for example when a sales manager requests a single 'target' figure for demand because he or she is unwilling or unable to cope with the challenge posed by a prediction expressed as a range of numbers, called an *interval forecast*. In fact the sales manager, whether he or she likes it or not, will typically have to face the potentially awkward questions raised by the twin, diametrically opposed risks involved in deciding how much stock to manufacture. Too much may result in high inventory costs, while too little may lead to unsatisfied demand and lost profits. Forecast users in other areas often face a similar quandary and so most forecasters do realize the importance of providing interval forecasts as well as (or instead of) point forecasts so as to enable users to

(1) Assess future uncertainty,

(2) Plan different strategies for the range of possible outcomes indicated by the interval forecast,

(3) Compare forecasts from different methods more thoroughly, and

(4) Explore different scenarios based on different assumptions more carefully.

Before proceeding further, we must define more carefully what is meant by an interval forecast. An *interval forecast* usually consists of an upper and a lower limit between which the future value is expected to lie with a prescribed probability. The limits are sometimes called *forecast limits* (Wei 1990) or *prediction bounds* (Brockwell & Davis 1991, p. 182), while the interval is sometimes called a *confidence interval* (Granger & Newbold 1986) or a *forecast region* (Hyndman 1995). I prefer the more widely-used term *prediction interval*, as used by Abraham & Ledolter (1983), Bowerman & O'Connell (1987), Chatfield (1996a), and Harvey (1989), both because it is more descriptive and because the term *confidence interval* is usually applied to interval estimates for fixed but unknown parameters. In contrast, a prediction interval (henceforth abbreviated P.I.) is an interval estimate for an (unknown) future value. As a future value can be regarded as a random variable at the time the forecast is made, a P.I. involves a different sort of probability statement from that implied by a confidence interval.

In this chapter, I restrict attention to computing a P.I. for a single observation at a single time horizon. I do not consider the more difficult problem of finding a simultaneous prediction region for a set of related future observations, either forecasts for a single variable at different horizons or forecasts for several variables at the same horizon. For example, it is common to want to forecast sales for each month of the coming year, say, and then find a 95% P.I. for each value independently of the rest. However, this tells us nothing about the overall probability that *at least* one future observation will lie outside its P.I. This combined probability will be (much) greater than five per cent and has to be evaluated using specialized techniques described by Lütkepohl (1991, Section 2.2.3) and Ravishankar, Wu & Glaz (1991).

NOTATION

An observed time series, containing *n* observations, is denoted by $x_1, x_2, \ldots, x_n$. Suppose we wish to forecast the value of the series *h* steps ahead. This means we want to forecast the observed value at time $(n+h)$. The integer *h* is called the *lead time* or *forecasting horizon* (*h* for horizon). The point forecast of the value at time $(n+h)$ made using the data up to time *n* is denoted by $\hat{x}_n(h)$. Note that it is essential to specify both the time at which a forecast is made *and* the forecasting horizon. When the observed value later becomes available, we can calculate the corresponding forecast error, denoted by $e_n(h)$, by

$$e_n(h) = x_{n+h} - \hat{x}_n(h). \qquad (1)$$

The notation for this forecast error, like that for the point forecast, specifies both the horizon and the time period when the forecast was made.

MODELS AND METHODS

Statisticians customarily regard the data as being observations on an underlying *model*, which is a *mathematical representation of reality* and is usually *approximate* rather than exact. In a model, the observation at time *t*, namely x_t, is regarded as being an observation on an underlying random variable, which is usually denoted by a capital letter, X_t, in contrast to the use of lower case letters for observed data. A typical model is the first-order *autoregressive* model, denoted by AR(1), for which

$$X_t = \alpha X_{t-1} + \varepsilon_t \qquad (2)$$

where α denotes a constant (with $|\alpha|<1$ for stationarity) and ε_t denotes the error at time *t*. More generally a model with additive errors can be represented by

$$X_t = \mu_t + \varepsilon_t \qquad (3)$$

where μ_t denotes the predictable part of the model. Engineers typically refer to μ_t as the *signal* and ε_t as the *noise*, and I think this terminology can be helpful because the 'error' term in the mathematical model is not really an error in the usual sense of the word. Statisticians sometimes refer to the error terms as the *innovations* or use the engineers terminology of *noise*. The signal in Equation (3) could, for example, include a linear trend with time and/or linear multiples of past values (called autoregressive terms) as in Equation (2). The noise could include measurement error and natural unpredictable variability. The $\{\varepsilon_t\}$ are usually assumed to be a sequence of independent normally distributed random variables with zero mean and constant variance σ_ε^2, which we write as $NID(0, \sigma_\varepsilon^2)$.

I draw a clear distinction between a forecasting *method* and a *model*. A *forecasting method* is a rule or formula for computing a point forecast from the observed data. As such, it is not a model, although it may be based on a model. For example, *exponential smoothing* is a method that computes a point forecast by forming a weighted average of the latest observation and the most recent point forecast. It can be shown that this method is optimal (meaning that it gives minimum mean-square error forecasts) for a particular type of model which can be written

$$X_t = X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1} \ , \qquad (4)$$

and which is customarily denoted as an ARIMA(0,1,1) model (Box, Jenkins & Reinsel 1994). Thus exponential smoothing is based on a model but is not a model itself.

There is a rich variety of forecasting methods, and the choice among them depends on many factors, such as background knowledge, the objectives, and the type of data. Given such a wide range of methods, it follows that a variety of approaches will be needed to compute P.I.s. It is helpful to categorize forecasting methods as (1) *univariate*, where $\hat{x}_n(h)$ depends only on past values of the given series, namely $x_n, x_{n-1}, \ldots.,$ (2) *multivariate*, where $\hat{x}_n(h)$ may also depend on other explanatory variables, and (3) *judgmental*. It can also be helpful to distinguish between *automatic* methods, requiring no human intervention, and *non-automatic* methods.

A further useful distinction is between methods that involve fitting an 'optimal' probability model and those that do not: the latter perhaps more familiar to the operational researcher and the former to the statistician when it is usually possible to compute theoretical P.I.s conditional on the fitted model. However, the practitioner with a large number of series to forecast may decide to use the same all-purpose procedure whatever the individual series look like, as for example when the Holt-Winters forecasting procedure is used for a group of series showing trend and seasonal variation. The method does not depend explicitly on any probability model, and no model-identification is involved. This means that forecasts need not be optimal for each individual series and it is not so easy to construct P.I.s.

In summary, a forecasting *method* (i.e., a rule for computing forecasts, such as exponential smoothing) may or may not be developed from a *model* (a mathematical representation of reality, such as an AR(1) model).

It is also useful to distinguish between the (observed) errors that arise from using a method and the (theoretical) errors which form part of a model. The forecast errors in Equation (1), namely $e_n(h)$, can be described as the observed *out-of-sample* forecast errors. They are not the same as the errors that form part of the mathematical representation of the model. For example, in Equations (2) and (3), the ε_t are theoretical error terms. It is also helpful to understand the distinction between the observed *out-of-sample* forecast errors (the $e_n(h)$) and the observed *within-sample* one-step-ahead 'forecasting' errors, namely $[\, x_t - \hat{x}_{t-1}(1) \,]$ for $t = 2, 3, \ldots, n$. When forecasts are obtained by fitting a model and computing minimum mean-square-error forecasts from the model, then the within-sample 'forecast' errors are the *residuals* from the fitted model, because they are the differences between the observed and the fitted values. Unlike the out-of-sample errors, they are not true ex-ante forecasting errors, because the model is typically determined by estimating parameters from all the data up to time *n*.

If one has found the correct model for the data, and if the model does not change, then one might expect the out-of-sample forecast errors to have properties similar to both the residuals and the true 'error' terms. In practice, these three types of error have rather different properties. First, the out-of-sample forecast errors may be calculated for different horizons, and it can be shown that they tend to get larger as the horizon gets longer for nearly all methods and models, because the errors at each time interval build up in a cumulative way. Thus it is only reasonable to compare the one-step-ahead out-of-sample forecast errors with the residuals. Second, the within-sample residuals and the one-step-ahead out-of-sample forecast errors both depend on *estimates* of the parameters used in the forecasting process, rather than on the true values. Because of this, it can be shown that, if a model has been fitted, then the (theoretical) error terms in the model will have properties different from both the (observed) residuals and the out-of-sample forecast errors. Third, the wrong forecasting method or model may be chosen or the underlying model may change, and this helps to explain why the out-of-sample forecast errors are typically found to have (much) larger variance than the residuals.

SOME PROBLEMS

Given their importance, it is regrettable that most companies do not regularly produce P.I.s for their internal forecasts (Dalrymple 1987), and that many economic predictions are still given as a single value (though my subjective impression is that this is slowly changing). Several reasons can be suggested for the infrequent provision of interval forecasts and for a lack of trust in their calibration properties when they are calculated:

(1) The topic has been rather neglected in the statistical literature. The authors of textbooks on time-series analysis and forecasting generally say surprisingly little about interval forecasts and give little guidance on how to compute them, except perhaps for regression and Box-Jenkins (ARIMA) models. Some relevant papers have appeared in statistical and forecasting journals, but they can be mathematically demanding, unhelpful, or even misleading or wrong. I focus on the *principles* for computing P.I.s. and include a summary of my earlier literature review (Chatfield 1993) as well as some more recent research, including work on the effects of model uncertainty on P.I.s (Chatfield 1996b).

(2) No generally accepted method exists for calculating P.I.s except for forecasts calculated conditional on a fitted probability model, for which the variance of forecast errors can be readily evaluated.

(3) Theoretical P.I.s are difficult or impossible to evaluate for many econometric models, especially multivariate models that contain many equations or that depend on non-linear relationships. In any case, when judgmental adjustment is used in the forecasting process (e.g., to forecast exogenous variables or to compensate for anticipated changes in external conditions), it is not clear how one should make corresponding adjustments to interval forecasts.

(4) Analysts sometimes choose a forecasting method for a group of series (e.g. in inventory control) by using domain knowledge and the common properties of the various series (e.g., seasonal or non-seasonal), with no attempt to find a probability model for each individual series. Then it is not clear if P.I.s should be based on the model, if any, for which the method is optimal. When a method is not based explicitly, or even implicitly, on a probability model, it is unclear how to proceed.

(5) Various ‘approximate’ procedures have been suggested for calculating P.I.s, but there are justified doubts as to their validity.

(6) Researchers have developed some alternative computational methods for calculating P.I.s, called *empirically based* and *resampling* methods, that do not rely on exact knowledge of the model, but their properties are not yet fully established and they have been little used in practice.

(7) Some software packages do not produce P.I.s at all, partly because of (1) to (4), while others produce them for regression and ARIMA models only or use 'approximate' formulae that are invalid.

(8) Empirical evidence suggests that P.I.s will tend to be too narrow on average, particularly for methods based on theoretical formulae, though less so for empirically based and resampling methods.

Given all these problems, it is clear that further advice and research are needed to clarify the situation.

SOME GENERAL PRINCIPLES FOR COMPUTING P.I.s

- **The importance of P.I.s:** *It is usually important to supplement point forecasts by computing interval forecasts*.

Three reasons were given in the Introduction to justify this principle, which some readers may find self-evident. Of particular importance is the general requirement to provide a measure of the uncertainty associated with any forecast. As corollaries, it follows that

(1) *Forecasters must have the skills to enable them to compute interval forecasts*, and

(2) *More attention should be given to providing the necessary methodology in the forecasting literature*.

- **The availability of theoretical formulae:** *Theoretical formulae are available for computing P.I.s for various classes of time-series model, including regression, ARIMA, and structural models, and also for some forecasting methods (as opposed to models), including various forms of exponential smoothing*.

This principle is the source of most P.I.s calculated in practice. The formulae are essentially of the same general form, namely that a 100(1-α)% P.I. for the value h steps ahead is given by

$$\hat{x}_n(h) \pm z_{\alpha/2}\sqrt{Var[e_n(h)]} \tag{5}$$

where appropriate formula for $\hat{x}_n(h)$ and for $Var[e_n(h)]$ are found for the method or model which is deemed appropriate and $z_{\alpha/2}$ denotes the appropriate (two-tailed) percentage point of a standard normal distribution.

The interval is symmetric about $\hat{x}_n(h)$, so that Equation (5) effectively assumes that the point forecast is unbiased. The usual statistic for assessing the uncertainty in forecasts of a *single* variable is the expected mean square prediction error (abbreviated PMSE), namely $E[e_n(h)^2]$ (but note that scale-independent statistics, such as the mean absolute prediction error (MAPE), will be preferred for *comparing* the accuracy of forecasts made for *different* variables, especially when measured on different scales (Armstrong & Collopy 1992). For an unbiased forecast, $E[e_n(h)^2] = Var[e_n(h)]$ so that the PMSE is equal to the latter expression. Forecasters generally assume unbiasedness (explicitly or implicitly) and work with Equation (5), which takes $Var[e_n(h)]$ as the PMSE. Thus, to apply Equation (5), the forecaster needs to be able to compute $Var[e_n(h)]$. Formulae are available for doing this for various classes of model, including regression, ARIMA, structural (state-space), and VARMA models, and Chatfield (1993, Section 4.2) gives the relevant references. However, theoretical formulae are not available for certain types of model, notably simultaneous-equation econometric models, especially when non-linearities are involved or when point forecasts are judgmentally adjusted. They are also not immediately available for forecasting *methods* that do not depend explicitly on a probability model (but see below).

In fact, the formulae for $Var[e_n(h)]$ typically given in the literature are what might be called 'true-model' PMSEs, because they assume that there is a true, known model and that the model parameters are known exactly. In practice, the parameters have to be estimated, and it is customary to substitute estimated values in the theoretical formulae. Does this matter? Chatfield (1993, Section 3) discusses this technical issue in detail. It can be shown that the effect of parameter uncertainty on the coverage of P.I.s gets smaller as the sample size gets larger (as would intuitively be expected; a mathematician would say that the effect is of order 1/*n*.). Moreover, this effect is likely to be of a smaller order of magnitude than some other effects, notably the effects of uncertainty about the structure of the model and the effects of errors and outliers, which I discuss later. However for sample sizes smaller than about 50, the effect of parameter uncertainty could be non-trivial, especially for models with many parameters used to predict at longer lead times. Nevertheless, given all other uncertainties, it is usually adequate to compute P.I.s using equation (5) by substituting parameter estimates into the true-model PMSE to get $Var\,[e_n(h)\,]$.

The above discussion concerns the use of theoretical formulae for $Var[e_n(h)]$ for various classes of model. A natural follow-up question is whether theoretical formulae can also be found for some forecasting *methods* (as opposed to models). As noted earlier, a forecasting method is sometimes selected without applying any formal model-identification procedure, although one should certainly choose a method appropriate to any trend or seasonality that is present. The question then arises as to whether P.I.s should be calculated by some computational procedure that does not depend on a model or by assuming that the method is optimal in the sense that the true model is the one for which the selected forecasting method is optimal.

For example exponential smoothing (ES) can be used for series showing no obvious trend or seasonality without necessarily trying to identify the underlying model. Now ES is known to be optimal for an ARIMA(0,1,1) model (Equation (4) above) and also for a particular structural (or state space) model, and both of these models lead to the same 'true-model' PMSE formula (Box, Jenkins & Reinsel 1994, p.153; Harrison 1967)

$$Var[e_n(h)] = [1 + (h{-}1)\alpha^2]\sigma_e^2 \qquad (6)$$

where α denotes the smoothing parameter and $\sigma_e^2 = Var[e_n(1)]$ denotes the variance of the one-step-ahead forecast errors. Should this formula then be used in conjunction with Equation (5) for ES even though a model has not been formally identified? I suggest that it is reasonable to use Equation (6) provided that the observed one-step-ahead forecast errors show no obvious autocorrelation and provided that no other obvious features of the data (e.g., trend) need to be modeled. However, there are some alternative P.I. formulae for ES that *should* be disregarded because they are based on inappropriate models (Chatfield 1993, Section 4.3).

It is possible to compute P.I.s for some methods without recourse to any model (Chatfield 1993, Section 4.2). If we assume that the method is optimal in the sense that the one-step-ahead errors are uncorrelated, then it may be possible to express $e_n(h)$ in terms of the intervening one-step-ahead errors and evaluate $Var[e_n(h)]$ in terms of $Var[e_n(1)]$. Then Equation (5) can still be used. Yar & Chatfield (1990) and Chatfield & Yar (1991) have applied this approach to the Holt-Winters method with additive and multiplicative seasonality respectively. The results in the multiplicative case are of particular interest because $Var\,[e_n(h)]$ does not necessarily increase monotonically with *h*. Rather P.I.s tend to be wider near a seasonal peak as might intuitively be expected. This sort of behavior is typical of non-linear models (Tong 1990, Chapter 6) and arises because forecasts from multiplicative Holt-Winters are not a linear combination of past observations.

The other obvious feature of Equation (5) is that it involves the percentage point of a standard normal distribution and so effectively assumes that the forecast errors are normally distributed. This leads on to an important corollary:

> *Checking normality: When using a symmetric P.I. that utilizes normal percentage points (as in Equation (5)), check that the normality assumption is at least approximately true.*

The analyst will typically be concerned about two main types of departure from normality in the distribution of the error terms. They are (a) *asymmetry* and (b) *heavy tails*. Heavy tails may be caused, for example, by occasional outliers, and this problem can be tackled by modifying Equation (5) by changing $z_{\alpha/2}$ to the appropriate percentage point of an alternative error distribution that is found either by using the empirical distribution of the residuals or by trying an alternative theoretical distribution with heavier tails than the normal. As regards asymmetry, some researchers have found evidence of its presence (Williams & Goodman 1971; Makridakis *et al.*, 1987). This is especially true (Armstrong & Collopy 1997) for annual economic variables that are non-negative (i.e., have a natural zero) and show steady growth so that it is the *percentage* change that is of particular interest. Then typically one finds that the residuals from an additive model fitted to the raw data are not symmetric but are skewed to the right.

Transformations. An asymmetric error distribution can usually be made more symmetric by transforming the data in some way, most often by taking logarithms. If a model is formulated for the logs and then used to compute point and interval forecasts for future values of the logged variable, then these will need to be transformed back to the original units to give forecasts of what is really required (Chatfield 1993, Section 4.8). Note that the so-called naive retransformed point forecast will not in general be unbiased. In other words, if the analyst takes logs of a variable, finds point forecasts of the logs and assumes they are unbiased, and then takes antilogs to get point forecasts of the original variable, then the latter forecasts will no longer be unbiased. It is possible to correct for this, but the correction is rarely used. Fortunately P.I.s have nicer properties under transformation in that the naive retransformed P.I. will have the correct prescribed probability. What does this mean? Suppose the analyst finds a 95% P.I. for the logarithm of the variable. If one takes antilogs of the upper and lower limits of this P.I. to get the retransformed P.I. for the original variable, then it can easily be shown that there will still be a 95% probability that this interval will include the future value of the original variable. This retransformed P.I. will generally be asymmetric, as it should be to reflect the asymmetry in the errors.

Non-linear models. The normality assumption also makes Equation (5) unsuitable for many non-linear models where it can be shown that the predictive distribution is generally not normal (e.g. Hyndman 1995). It could for example have two peaks (i.e. be bimodal). In extreme cases, a sensible P.I. could even comprise two (or more) disjoint intervals and then the term *forecast region* seems more appropriate than P.I. Unfortunately it can be difficult to evaluate conditional expectations more than one step ahead for non-linear models. Moreover the width of P.I.s need not necessarily increase with lead time for such models. This means that there may be no alternative to evaluating the complete predictive distribution (i.e., the complete distribution of future values that might result) at different lead times for a non-linear model even though this may be computationally demanding.

Conditional P.I.s. A more subtle point is that, even for a *linear* model with normally distributed errors, the one-step-ahead forecast error distribution, *conditional on the latest value*, will not in general be exactly normal when model parameters are estimated from the same data used to compute forecasts (Chatfield 1993, Section 4.1). The correction to the normal approximation for linear models seems likely to be of a smaller order of magnitude in general than other corrections, although some authors (e.g., Harvey 1989, p.32) do suggest replacing $z_{\alpha/2}$ in Equation (5) by the appropriate percentage point of a t-distribution when model parameters are estimated. However, this is not based on general theory and in any case makes little difference except for very short series (e.g., less than about 20 observations) where other effects (e.g., model and parameter uncertainty) are likely to be more serious anyway.

For non-linear models, such as GARCH models, the difference between conditional and unconditional P.I.s can be much more substantial, and Christoffersen (1998) has proposed a framework for assessing conditional forecast evaluation. The basic idea is that P.I.s should be relatively narrow in times of stability but wider when behavior is more volatile.

Summary. Equation (5) is widely used for computing P.I.s for various models and methods, but should preferably be used only after checking that the underlying assumptions, especially normality, are at least reasonably valid.

- **Beware so-called approximate formulae:** *It is generally unwise to base P.I.s on one of the various so-called approximate formulae that have been suggested for calculating Var* $[e_n(h)]$.

When theoretical formulae are not available, (and even when they are), some writers have suggested a variety of simplistic 'approximate' formulae for calculating $Var\,[e_n(h)]$ for use with Equation (5). This is unfortunate given that the approximations are often (very) poor as Chatfield (1993, Section 4.4) demonstrates. The best known example is the general 'approximate' formula that

$$Var\,[e_n(h)] = h\,\sigma_e^2 \qquad (7)$$

where $\sigma_e^2 = Var\,[e_n(1)]$ denotes the variance of the one-step-ahead forecast errors. In fact Equation (7) is true only for a random walk model; for other methods and models it can be seriously in error and should not be used. When theoretical formulae are not available, it will still usually be possible to use empirically based or resampling methods and so there is no real reason why the 'approximate' formulae should ever be used.

- **Availability of computational alternatives:** *When using a model of doubtful validity or for which the theoretical PMSE formula is not available, be aware that alternative computationally intensive approaches to the construction of P.I.s are available. They include: (1) empirically based P.I.s that rely on the properties of the observed distribution of residuals (rather than on an assumption that the model is true), and (2) simulation and resampling methods, which involve generating possible future paths for a series, either by simulating future random variables from the fitted model or by resampling the distribution of 'errors' in some way.*

These methods generally require fewer (or even no) assumptions about the underlying model, have much promise, and are starting to be used.

Chatfield (1993, Section 4.5) reviews the use of empirically based P.I.s. The simplest type of procedure involves applying the forecasting method to past data, finding the within-sample 'forecast' errors (i.e., the residuals) at 1, 2, 3, ... steps ahead for forecasts made from all available time origins in the period of fit, and then finding the variance of these errors at each lead time. Let $s_{e,h}$ denote the standard deviation of the h-steps-ahead errors. Then an approximate empirical $100\,(1-\alpha)\%$ P.I. for X_{n+h} is given by $\hat{x}_n(h) \pm z_{\alpha/2}\, s_{e,h}$. The approach often works reasonably well and gives results comparable to theoretical formulae when the latter are available. However, the values of $s_{e,h}$ tend to be unreliable, especially for small n and large h, and, even with a reasonably long series, one may find that the values do not increase monotonically with *h*. Thus it may be wise to smooth the values in some way, perhaps by averaging them over adjacent values of *h*, though I am not aware that advice on this has actually appeared in print. Another problem is that the values of $s_{e,h}$ are based on model-fitting errors rather than on post-sample forecast errors. There is empirical evidence that the characteristics of the distributions of these two types of error are generally not the same. In particular, out-of-sample forecast errors tend to have larger variance (e.g., Makridakis & Winkler 1989; Chatfield 1996b). Thus P.I.s produced in this way tend to be too narrow (as are theoretical formulae).

In an earlier related proposal, Williams & Goodman (1971) suggested splitting the past data into two parts, fitting the method or model to the first part and making predictions about the second part. The resulting prediction errors are more like true forecast errors. One then refits the model with one additional observation in the first part and one less in the second part; and so on. For some monthly data on numbers of business telephone lines in service, Williams and Goodman found that the distribution of forecast errors tended to approximate a gamma distribution rather than a normal distribution. They constructed P.I.s using the percentage points of the empirical distribution, thereby avoiding any distributional assumptions, and obtained promising results. The method has been little used in practice, presumably because the heavy computational demands were beyond the resources of the early 1970s, but is now due for reassessment.

Simulation and resampling methods provide an alternative to empirically based P.I.s. Given a probability time-series model, it is possible to *simulate* both past and future behavior by generating an appropriate series of random error terms from some assumed parametric distribution (e.g. a normal distribution) and hence constructing a sequence of possible past and future values. This process can be repeated many times, and this makes it possible to evaluate P.I.s at different horizons by simply finding the interval within which the required percentage of simulated future values lies. Alternatively, instead of sampling the errors from an assumed parametric distribution, it is possible to sample from the empirical distribution of past residuals (the fitted errors). This is called *resampling* (or *bootstrapping* in the statistical literature) and is a distribution-free approach. Again the idea is to generate a sequence of possible future values and find appropriate P.I.s by inspection. Chatfield (1993, Section 4.6) reviews the literature in this area. Veall (1989) suggests that resampling methods are particularly helpful in dealing with the shortcomings of asymptotic and analytic approaches in econometrics, especially when models are very complex, or non-linear or data sets are small.

Statisticians generally use the term *bootstrapping* in quite a different way from that used by judgmental researchers, to describe the process of taking a random sample of size *n* from a set of independent observations of size *n* where observations are taken *with replacement*. This means that some observations from the original recorded sample will occur more than once in the bootstrap sample and some not at all. In a time-series context, this type of sampling would make no sense because the observations are not independent but are ordered through time. This explains why statisticians usually bootstrap time-series data by resampling the fitted errors (which are hopefully close to independence) rather than the actual observations, but this does not disguise the fact that it is generally more difficult to resample correlated data, such as time series, than to resample independent observations. Furthermore, resampling fitted errors makes the procedure more dependent on the fitted model. Several writers (e.g., Thombs & Schucany 1990) give much more information as to how to carry out resampling for time-series data and I do not give details here. McCullough (1994; 1996) describes some recent work on bootstrapping autoregressive and multiple regression models. While it is very much an ‘in’ method, bootstrapping does not always work. Sadly practitioners tend to suppress poor results when they happen. Meade & Islam (1995) report one example where bootstrapping gave poor results in regard to finding P.I.s for growth curve models. This is a tricky problem, largely neglected in the literature, because a model such as a Gompertz curve is non-linear in the parameters and in addition it is not obvious how to specify the error structure. Meade & Islam (1995, especially p. 427) investigate three possible methods for computing growth curve P.I.s and find those based on bootstrapping are "far too narrow".

- **Consider a Bayesian approach:** *A Bayesian approach may make it possible to find the complete predictive distribution for a future value and hence compute Bayesian interval forecasts. The Bayesian approach may also make it feasible to use a mixture of models, rather than a single model.*

Bayesian methods have been attractive to some statisticians for many years because of the philosophical coherence of the general approach, but they have often proved difficult or impossible to implement in practice. However, recent advances in computational methods have meant that many problems can now be solved with a Bayesian approach, albeit with quite extensive numerical work in most cases. In forecasting, the Bayesian multi-period ahead predictive density does not have a convenient closed form for many models, and so Bayesian statisticians will need to consider alternatives. Some sort of approximation may be possible to compute interval forecasts (Thompson

& Miller 1986, Section 3), or it may be possible to simulate the predictive distribution rather than try to obtain or approximate its analytic form. The phrase ‘Bayesian forecasting’ is often used to describe a particular approach based on a class of models called *dynamic linear models* (West & Harrison 1997). Chatfield (1993, Section 4.7) gives a brief review of the literature up to 1993.

A Bayesian approach may also seem natural when the analyst decides to rely, not on a single ‘best’ model (which may be wrongly identified or may change through time), but on a *mixture* of models. It is well known that combining forecasts from different sources generally gives more accurate point forecasts on average (Clemen 1989) than any single point forecast. Unfortunately there is no simple analytic way of computing the corresponding P.I.s to go with a combined forecast of this type, although it may be possible to use some sort of resampling method. However, a Bayesian formulation may enable the analyst to compute P.I.s for a combined forecast from a set of models that appear to be plausible for a given set of data. To do this, one uses a technique called *Bayesian Model Averaging* (Draper 1995) which is too large a topic to cover here. Draper’s (1995) Example 6.1 is particularly instructive in motivating the use of model averaging by demonstrating that conditioning on a single model can seriously underestimate the effect of model uncertainty. He assessed 10 possible econometric models that were proposed for predicting the price of oil from data up to 1980. The point and interval forecasts of the price in 1990 produced by the different models were often very different, but none of the intervals included the actual value which resulted. A model uncertainty audit suggested that only about 20% of the overall predictive variance could be attributed to uncertainty about the future conditional on the selected model and yet that is normally the only uncertainty that the analyst takes into account.

Although computational advances have been impressive, Bayesian methods are still not easy to implement. Recently analysts have begun to explore the use of a complex general-purpose simulation tool called *Markov Chain Monte Carlo* (abbreviated MCMC or MC^2) methods (e.g., Barnett, Kohn & Sheather, 1996; 1997) and the use of MCMC may enable the analyst to select a model, estimate parameters, and detect outliers all at the same time, yielding P.I.s that allow for model uncertainty and parameter estimation error. I have no practical experience with this procedure and will not attempt to comment on its potential.

- **Judgmental P.I.s:** *Judgment may be used to produce P.I.s, but empirical evidence suggests that they will generally be too narrow.*

Generally speaking, analysts are overconfident about their ability in judgmental forecasting and in behavioral decision theory (Armstrong 1985, pp. 138-145; O'Connor & Lawrence, 1989; 1992). Recently Wright, Lawrence & Collopy (1996) summarized past empirical findings by saying that "the evidence on the accuracy and calibration of judgmental P.I.s is not very encouraging". This is disappointing. Because the topic is outside the scope of this chapter with its quantitative emphasis, I will not pursue the topic here, but refer the reader to Arkes (1998).

Choosing a method to compute P.I.s: Choosing an appropriate method for computing P.I.s may appear difficult after reading about the many different possible approaches. In practice, the choice is often determined by the choice of forecasting method, which depends in turn on such factors as the objectives and type of data.

Theoretical P.I. formulae are available for many models. When the analyst chooses a forecasting method based on a particular model, the theoretical formulae are easy to implement and are widely used. However, such formulae are not available for some complex or non-linear models. Moreover the formulae are appropriate only if the fitted model is correctly identified, and the possibility that the model may be misspecified or may change in the forecast period is a serious problem. This is why it is essential to carry out appropriate diagnostic checks on the model, for example, to check that the residuals (the one-step-ahead forecast errors) are approximately uncorrelated.

When there are many series to forecast, the analyst usually chooses a simple automatic *method* and will then also need a simple method for computing P.I.s. Formulae based on the assumption that the method is optimal are widely used, but, as for model-based procedures, it is important to carry out appropriate diagnostic checks to make sure that the method really is sensible.

When a forecasting method or model is chosen for which the PMSE is not available or for which there are doubts about the underlying assumptions (if any), it may be necessary to use empirically based or resampling methods, which are nearly always available and which require fewer assumptions. They can be computationally demanding (especially resampling) but have great promise, and should arguably be used more than they are.

- **P.I.s are generally too narrow on average**.

In practice, analysts typically find, for example, that more than five percent of future observations will fall outside 95% P.I.s on average, especially when calculated using Equation (5) in genuine out-of-sample mode. Chatfield (1993, Section 5) reviews the empirical evidence for this and suggests the following possible reasons, not all of which need apply in any particular situation. They include

(1) The error distribution may not be normal. It may be asymmetric or heavy-tailed (perhaps due to occasional outliers); there may also be errors in the data that will contaminate the apparent 'error' distribution;

(2) Multivariate forecasts may require forecasts of exogenous variables;

(3) The 'true' model (if one exists) may not have been identified correctly;

(4) Even when the true model is correctly identified, the model parameters are unknown and have to be estimated;

(5) The underlying model may change through time, during the period of fit or in the future.

I discussed problem (1) earlier in regard to the 'Checking Normality' corollary. If non-normality is present, one can use an alternative parametric distribution for the errors or rely on the empirical distribution of the residuals. Outliers and errors will not only affect the perceived error distribution but also complicate model identification. Moreover, when an outlier is near the forecast origin, it is well known that it can have a disproportionate effect on point forecasts and on associated P.I.s (Ledolter 1989),

Problem (2) partly explains why multivariate forecasts need not be as accurate as univariate forecasts, contrary to many people's intuition (Ashley 1988).

As regards problem (3), it is always tempting to search for the 'true' model by (over)fitting the data with more and more complicated models to improve the fit. However, empirical evidence suggests that more complicated models, which give a better fit, do not necessarily give better out-of-sample forecasts. This has certainly been my experience using Box-Jenkins and neural network models (Faraway & Chatfield, 1998). The analyst effectively admits ignorance as to what the 'true' model is when he/she searches for the best-fitting model over what may be a wide class of models. It is therefore illogical that analysts then typically ignore model uncertainty and make forecasts as if the fitted model were known to be true in the first place (Chatfield 1996b). It is well known, for example, that (a) least-squares theory does not apply when the *same data are used to both formulate and fit a model* as typically happens in time-series analysis, and (b) when a model has been selected as the best-fitting model, the resulting parameter estimates will be biased and the fit will appear to be better than it really is. Picard & Cook (1984) call this the optimism principle.

When formulating a model, the use of appropriate diagnostic checks seems likely to lead to a fitted model that is at least a good approximation. Model checking is an integral part of the Box-Jenkins model-identification process (Box, Jenkins & Reinsel 1994, Chapter 8) and has come to be part of time-series modelling more generally. Even when using a forecasting method that does not depend explicitly on a probability model, one should still make checks on the one-step-ahead forecast errors to ensure, for example, that they are approximately uncorrelated.

Problem (4) can sometimes be dealt with by using PMSE formulae incorporating correction terms for parameter uncertainty. However the corrections are typically of order $1/n$ and of less importance than other factors (except perhaps for short series).

As regards (5), a model may change through time either because of a slowly changing structure or because of a sudden shift or turning point, such as the sudden changes to many economic variables that resulted from the 1973 oil crisis and the 1990 Gulf war. The prediction of change points is a topic of much current interest. It is notoriously difficult to do; Makridakis (1988, p. 479) asserts that "empirical evidence has shown that predicting cyclical turning points is extremely difficult or impossible".

These reasons help to explain why post-sample forecast errors tend to have larger variance than model-fitting errors as found empirically, for example, by Makridakis & Winkler (1989). Chatfield (1996b, Example 2) provides a recent demonstration adapted from the results of Faraway & Chatfield (1998), who fitted various neural networks to a set of data usually called the airline data. They found that the standard deviation of the one-step-ahead prediction errors in the test set (out-of-sample) was typically about twice the corresponding value in the training set (the fit sample), but this ratio was even larger for more complicated models (with more parameters) which gave a better fit but poorer out-of-sample performance.

Various modifications to Equation (5) have been suggested so as to make P.I.s realistically wide (Gardner 1988). However, for a 95% probability, they may become so embarassingly wide that they are of little practical use other than to indicate the high degree of future uncertainty. Granger (1996) suggest using 50%, rather than 95%, P.I.s because this gives intervals that are better calibrated in regard to their robustness to outliers and to departures from model assumptions. Such intervals will be narrower but imply that a future value has only a 50% chance of lying inside the interval. This seems undesirable. So what should be done?

Despite the above problems, I generally prefer, on grounds of simplicity, to use a theoretical formula that incorporates a normality assumption, as in Equation (5), provided that such a formula is available. As a compromise, I would use 90% (or perhaps 80%) intervals, rather than 95% (or 50%) intervals, to avoid ‘tail’ problems. When a series is reasonably well-behaved, this approach seems to work well enough. I also recommend stating explicitly that the use of Equation (5) assumes (1) the future is like the past with all the dangers this entails, and (2) the errors are approximately symmetric (if not, then a log transformation may be necessary). Alternative approaches may give somewhat better calibration in general but are generally much more complicated and not necessarily worth the extra effort. Major problems with Equation (5) generally arise because of a sudden change in the underlying structure, and then no method of computing point or interval forecasts is likely to be successful.

Whatever checks are made and whatever precautions are taken, it is still impossible to be certain that one has fitted the correct model or to rule out the possibility of structural change in the present or the future. Chatfield (1993, Section 7) gives one example that illustrates the overriding importance of good model identification. In this example, the point forecasts for the variable being analysed were generally poor because of a large, and perhaps unforeseeable, increase towards the end of the data. Two models were fitted to the same data. Both were plausible in terms of their fit. However the P.I.s for the non-stationary ARIMA(1,1,0) process were much wider than those for the alternative stationary AR(2) process. Analysts sometimes see wide P.I.s as indicating ‘failure’, either to fit the right model or to get a usable interval, but here the wider P.I.s resulting from the non-stationary process were more realistic in allowing for higher uncertainty. Clearly getting a narrower interval is not necessarily better. The difference between the widths of the P.I.s from the two models is much larger than that resulting from parameter uncertainty, and helps emphasize the special importance of model identification, particularly in regard to deciding whether the data are stationary or not.

Given the difficulty of identifying the ‘true’ model, even if there is one, the analyst should consider using a mixture of models, rather than a single model, or use a forecasting method that is not model based but is deliberately designed to be adaptive and robust. Researchers have done much work on such methods, exemplified by some successful results using the Kalman filtering approach based on state-space models.

IMPLICATIONS FOR PRACTITIONERS

The computation of interval forecasts can be of vital importance in planning and decision making. A variety of approaches to computing P.I.s are available, and I give some general principles to guide the practitioner in deciding *which* approach to use and *how*.

A theoretically satisfying way of computing P.I.s is to formulate a model that provides an adequate approximation to the given time series data, to evaluate the resulting prediction mean square error (PMSE), and then to use Equation (5). Although it may be possible to incorporate a correction term in the PMSE to allow for parameter uncertainty, this is usually of order $1/n$ and is often small compared with other uncertainties. Thus it is usually omitted (rightly or wrongly). By using a theoretical formula based on a model, one assumes that there is a true model and that it has been correctly identified. This identification must be correct not only in regard to the primary structure of the model, as for example which lagged variables are to be incorporated in an autoregressive model, but also in regard to the (secondary) error assumptions, as for example that the errors are normally distributed. When theoretical formulae are not available or there are doubts about model assumptions, the use of empirically based or resampling methods should be considered as a general-purpose alternative.

The practitioner should bear in mind the distinction between a forecasting *method* (an algorithm for computing a forecast) and a forecasting *model* (a mathematical representation of reality). A method may or may not depend explicitly or implicitly on a model. Thus for large groups of series, practitioners sometimes choose a forecasting method to use with all the series in the group. Then, for simplicity, P.I. formulae are usually based on the model for which the method is optimal, but the decision to do so should be supported by carrying out appropriate checks on the one-step-ahead forecasting errors, for example, to ensure that they are approximately uncorrelated.

Perhaps my main message in this chapter is that the analyst should normally compute P.I.s, but that he or she should not trust the results blindly. P.I.s tend to be too narrow in practice for a variety of reasons, not all of which can be foreseen. There is no general method for dealing with this. I prefer to compute P.I.s based on the usual assumptions but to spell out these assumptions clearly for the forecast user. For example, I would explicitly state that errors are assumed to be normally distributed and that the fitted model has been identified correctly. As such assumptions are hard to verify or may not be true, *all comparisons of forecasting methods and models should be made on the basis of out-of-sample forecasts rather than on measures of fit.*

IMPLICATIONS FOR RESEARCHERS

We need more research on empirically based and resampling methods to give theoretical and practical guidance to forecasters. In particular, for an empirically based approach, we need to find methods for smoothing the values of the *h*-steps-ahead forecast error standard deviations, $s_{e,h}$. We also need clearer guidance on how to bootstrap (correlated) time series data.

Given that P.I.s are generally too narrow, we need more empirical evidence to see how this effect is related to the type of data (monthly, quarterly or annual) and to the context (e.g., presence or absence of domain knowledge). We need more research to see how P.I.s constructed conditional on a 'best-fit' model can be widened to allow for model uncertainty. Out-of-sample forecasting accuracy is typically much worse than in-sample fit, and we need more empirical evidence to describe such differences. At the same time, we need some general theoretical guidance on the effects of model uncertainty if possible.

We need more empirical guidance on the form of the distribution of errors to see what error assumptions are sensible in general and when appropriate action may be needed to cope with non-normality. For example, it would be helpful to know what sort of data are typically non-normal and whether the resulting problems can be overcome by taking logs of the data.

Finally, we need to investigate further the possibility of using a mixture of models, perhaps via Bayesian model averaging, rather than relying on a single model.

REFERENCES

Abraham, B. & J. Ledolter (1983), *Statistical Methods for Forecasting*. New York: Wiley.

Arkes, H. (1998), "Overconfidence in judgmental forecasting," In this volume.

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

Armstrong, J.S. & F. Collopy (1992), "Error measures for generalizing about forecasting methods: Empirical comparisons," *International Journal of Forecasting*, **8**, 69-80.

Armstrong, J.S. & F. Collopy (1997), "Prediction intervals for extrapolation of annual economic data: Evidence on asymmetry corrections" (in preparation).

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

Barnett, G., R. Kohn, & S.J. Sheather (1996), "Robust estimation of an autoregressive model using Markov chain Monte Carlo," *Journal of Econometrics*, **74**, 237-254.

Barnett, G., R. Kohn, & S.J. Sheather (1997), "Robust Bayesian estimation of autoregressive-moving average models," *Journal of Time Series Analysis*, **18**, 11-28.

Bowerman, B.L. & R.T. O’Connell (1987), *Time Series Forecasting*, 2nd edn. Boston: Duxbury Press.

Box, G.E.P., G.M. Jenkins, & G.C. Reinsel (1994), *Time-Series Analysis, Forecasting and Control*, 3rd edn. San Francisco: Holden-Day.

Brockwell, P.J. & R.A. Davis (1991), *Time Series: Theory and Methods*, 2nd edn. New York: Springer-Verlag.

Chatfield, C. (1993) "Calculating interval forecasts" (with discussion), *Journal of Business and Economic Statistics*, **11**, 121-144.

Chatfield, C. (1996a), *The Analysis of Time Series*, 5th edn. London: Chapman and Hall.

Chatfield, C. (1996b), "Model uncertainty and forecast accuracy," *Journal of Forecasting*, **15**, 495-508.

Chatfield, C. & M. Yar (1991), "Prediction intervals for multiplicative Holt-Winters," *International Journal of Forecasting*, **7**, 31-37.

Christoffersen, P.F. (1998) "Evaluating interval forecasts", *International Economic Review* (forthcoming).

Clemen, R.T. (1989), "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, **5**, 559-583.

Dalrymple, D.J. (1987), "Sales forecasting practices: Results from a United States survey," *International Journal of Forecasting*, **3**, 379-391.

Draper, D. (1995), "Assessment and propagation of model uncertainty" (with discussion), *Journal of the Royal Statistical Society, Series B*, **57**, 45-97.

Faraway, J. & C. Chatfield (1998), "Time-series forecasting with neural networks: A comparative study using the airline data," *Applied Statistics*, **47**, 231-250.

Gardner, E.S. (1988), "A simple method of computing prediction intervals for time series forecasts," *Management Science*, **34**, 541-546.

Granger, C.W.J. (1996), "Can we improve the perceived quality of economic forecasts?" *Journal of Applied Econometrics*, **11**, 455-473.

Granger, C.W.J. & P. Newbold (1986), *Forecasting Economic Time Series*, 2nd edn. New York: Academic Press.

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

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: C.U.P.

Hyndman, R.J. (1995), "Highest-density forecast regions for non-linear and non-normal time series models," *Journal of Forecasting*, **14**, 431-441.

Ledolter, J. (1989), "The effect of additive outliers on the forecasts from ARIMA models," *International Journal of Forecasting*, **5**, 231-240.

Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Makridakis, S. (1988), "Metaforecasting," *International Journal of Forecasting*, **4**, 467-491.

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

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

McCullough, B.D. (1994), "Bootstrapping forecast intervals: An application to AR(*p*) models," *Journal of Forecasting*, **13**, 51-66.

McCullough, B.D. (1996), "Consistent forecast intervals when the forecast-period exogenous variables are stochastic," *Journal of Forecasting*, **15**, 293-304.

Meade, N. & T. Islam (1995), "Prediction intervals for growth curve forecasts," *Journal of Forecasting*, **14**, 413-430.

O'Connor, M. & M. Lawrence (1989), "An examination of the accuracy of judgemental confidence intervals in time series forecasting," *Journal of Forecasting*, **8**, 141-155.

O'Connor, M. & M. Lawrence (1992), "Time series characteristics and the widths of judgemental confidence intervals," *International Journal of Forecasting*, **7**, 413-420.

Picard, R.R. & R.D. Cook (1984), "Cross-validation of regression models," *Journal of the American Statistical Association*, **79**, 575-583.

Ravishankar, N., L. S-Y. Wu, & J. Glaz (1991), "Multiple prediction intervals for time series: comparison of simultaneous and marginal intervals," *Journal of Forecasting*, **10**, 445-463.

Thombs, L.A. & W.R. Schucany (1990), "Bootstrap prediction intervals for autoregression," *Journal of the American Statistical Association*, **85**, 486-492.

Thompson, P.A. & R.B. Miller (1986), "Sampling the future: A Bayesian approach to forecasting from univariate time series models," *Journal of Business & Economic Statistics*, **4**, 427-436.

Tong, H. (1990), *Non-Linear Time Series*. Oxford: Clarendon Press.

Veall, M.R. (1989), "Applications of computationally-intensive methods to econometrics," *Bulletin of the I.S.I.*, 47th Session, 75-88.

Wei, W.W.S. (1990), *Time Series Analysis*. Redwood City, Cal.: Addison-Wesley.

West, M. & J. Harrison (1997), *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer-Verlag.

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

Wright, G., M.J. Lawrence, & F. Collopy (1996), "Editorial: The role and validity of judgment in forecasting," *International Journal of Forecasting*, **12**, 1-8.

Yar, M. & C. Chatfield (1990), "Prediction intervals for the Holt-Winters forecasting procedure," *International Journal of Forecasting*, **6**, 1-11.

Acknowledgements: I thank Scott Armstrong and several referees for constructive suggestions.

Full address: Department of Mathematical Sciences, University of Bath, Bath, Avon, BA2 7AY, U.K.

Fax: U.K.-1225 826492

Email: cc@maths.bath.ac.uk