MANAGEMENT SCIENCE
Vol. 34, No. 4, April 1988
Printed in U.S.A.

A SIMPLE METHOD OF COMPUTING PREDICTION INTERVALS FOR TIME SERIES FORECASTS*

EVERETTE S. GARDNER, JR.
College of Business Administration, University of Houston, Houston, Texas 77004

Theoretical approaches to computing prediction intervals require strong assumptions that do not appear to hold in practice. This paper presents an empirical approach to prediction intervals that assumes very little. During model-fitting, variances of the errors are computed at different forecast leadtimes. Using these variances, the Chebyshev inequality is applied to determine prediction intervals. Empirical evidence is presented to show that this approach gives reasonable results. For example, using the 111 series in the *M*-competition, 95% prediction intervals actually contain 95.8% of post-sample observations.
(FORECASTING—TIME SERIES)

1. Introduction

Almost all point forecasts are wrong. Thus prediction intervals are needed to indicate the likely precision of the forecasts for management planning. Prediction intervals are especially helpful in forecasting for inventory control, where safety stocks depend on the probability distribution of leadtime demand. Another application of prediction intervals is to identify outliers in time series.

Traditionally, prediction intervals are computed by making one of two critical assumptions: (1) the correct model has been identified or (2) the generating process for the time series is known. In most cases, these assumptions make it straightforward to derive closed-form expressions for variances of the forecast errors at different leadtimes.

For example, Box and Jenkins (1976) rely on the first assumption to obtain prediction intervals for ARMA models. Advocates of exponential smoothing do not assume that the correct model is used. Instead they assume that the generating process is known and derive variances based on the relationship between this process and the model at hand. See, for example, the expressions for variances of exponential smoothing models in Brown (1963), McKenzie (1976, 1984, 1986), and Sweet (1985). To simplify matters, both ARMA modelers and exponential smoothers almost always make another critical assumption, that the errors are normally distributed.

One problem with this theoretical work on prediction intervals is that it is impossible to obtain closed-form expressions for the variances of nonlinear forecasting systems. For example, there is no closed form for the popular Holt-Winters class of exponential smoothing systems with linear trend and multiplicative seasonality (McKenzie 1984). For such systems, an empirical method is the only alternative for computing prediction intervals.

Another problem is that theoretical prediction intervals often yield poor results in practice. In a reexamination of the 111 time series from the *M*-competition (Makridakis et al. 1982), Lusk and Belhadjali (1986) found that 95% prediction intervals for ARMA models contained only about 80% of post-sample observations. Makridakis and Hibon (1986) obtained similar results for other time series models.

One solution to these problems is the approach taken by Williams and Goodman (1971), who developed empirical distributions of post-sample forecast errors. Their procedure is to fit a model to a sample of a time series, make a set of forecasts from one

* Accepted by Vijay Mahajan; received January 12, 1987. This paper has been with the author 1 month for 1 revision.

0025-1909/88/3404/0541$01.25

time origin, compile the post-sample errors, and refit the model with one additional observation. These steps are repeated until the end of the data is reached. The Williams and Goodman procedure is so tedious that no applications since 1971 appear to have been reported.

This paper suggests a simple method of computing prediction intervals that avoids the need to make assumptions about the validity of the model, the form of the generating function, or the form of the distribution of forecast errors. Although the method is empirical, it is far less tedious than the Williams and Goodman approach.

2. The Method

Given that a forecasting model has been fitted to a time series, the first step is to compute variances of fitted errors at different leadtimes. For example, suppose we have a model selected on the usual basis of one-step-ahead fit. One pass through the data is made to compute the variance of the fitted errors at one-step-ahead. A second pass is made to compute the variance at two-steps-ahead. It is important to understand that the forecasting model is not re-estimated. We simply make two-step-ahead forecasts with the same model and compute the variance of the fitted errors. This process is continued until an individual variance is computed for each desired leadtime. For example, if prediction intervals are desired for leadtimes 1–12, there will be 12 individual variances. The advantage of this method of computing variances is that the validity of the model and the form of the generating function are irrelevant. We simply record the performance of the model.

The second step in the procedure is to compute standard errors at each leadtime. The final step is to apply a multiplier to each standard error that yields the desired prediction intervals. The multiplier is based on the Chebyshev inequality (Wilks 1962), which sets a bound on the amount of probability within given limits for any distribution with finite variance. Since the Chebyshev inequality is well known, only the main result is given here. Let Y be a random variable with mean μ and variance σ^2. The inequality is

$$P[|(Y-\mu)/\sigma| \geqslant \epsilon] \leqslant 1/\epsilon^2. \tag{1}$$

No matter what the actual distribution, equation (1) states that the probability of an observation falling beyond ϵ standard errors from the mean is at most $1/\epsilon^2$.

Admittedly the Chebyshev inequality generates crude bounds for many known distributions. For example, if the true distribution of errors is normal, the Chebyshev bounds are much too wide. But in practice the true distribution of errors is never known. Makridakis and Winkler (1985) were unable to find any distribution that gave a reasonable fit to the errors in a collection of 1,001 time series. Their study suggests that it is dangerous to assume some arbitrary distribution for the errors. The simple expression in (1) is at least a starting point for the development of prediction intervals.

3. Empirical Results

The time series from the M-competition (Makridakis et al. 1982) were used to test this method of computing prediction intervals. Forecasts were produced using the class of exponential smoothing systems with damped trends developed by Gardner and McKenzie (1985). Other computational details sufficient to replicate the results below are available in Gardner (1986).

Forecasting was done in exactly the same manner as Makridakis. Let the length of a series be N. For annual series, the forecasting model was fitted to the first $N-6$ observations. A set of 6 forecasts (for leadtimes 1–6) was made at time origin $N-6$. For quarterly data, 8 forecasts were made at $N-8$. For monthly data, 18 forecasts were made at $N-18$.

FIGURE 1. Percentage of Post-Sample Observations Inside Prediction Intervals: 111 Time Series.

Forecast Leadtime	90% Target		95% Target	
	Chebyshev	Normal	Chebyshev	Normal
1	94.6%	82.9%	97.3%	88.3%
2	93.7	77.5	99.1	83.8
3	94.6	73.0	96.4	85.6
4	91.9	71.2	95.5	80.2
5	84.7	72.1	93.7	77.5
6	86.5	66.7	91.9	72.1
7	91.2	70.3	95.6	73.6
8	92.3	69.2	96.7	74.7
9	92.6	76.5	98.5	80.9
10	89.7	73.5	98.5	75.0
11	94.1	73.5	97.1	77.9
12	89.7	72.1	95.6	79.4
13	86.8	70.6	94.1	73.5
14	91.2	73.5	97.1	79.4
15	89.7	66.2	94.1	73.5
16	92.6	79.4	95.6	83.8
17	88.2	75.0	94.1	82.4
18	89.7	67.6	94.1	76.5
All	90.8%	72.9%	95.8%	79.1%

The Makridakis data include 1,001 time series. Figure 1 gives results for a sample of 111 series taken from the 1,001. This is the same sample discussed in Makridakis et al. (1982). For each series, 90% and 95% prediction intervals were computed and the number of post-sample observations inside the intervals was recorded. The Chebyshev prediction intervals contain approximately the desired percentages: 90% intervals contain 90.8% of post-sample observations, while 95% intervals contain 95.8%.

Figure 1 also gives results for prediction intervals assuming the normal distribution and using the method of computing variances described above. Normal 90% intervals contain only 72.9% of post-sample observations, while normal 95% intervals contain only 79.1%. The normal percentages inside prediction intervals are comparable to the results reported by Williams and Goodman (1971), Lusk and Belhadjali (1986), and Makridakis and Hibon (1986).

Figure 2 gives results for the complete collection of 1,001 time series. Percentages inside the prediction intervals are similar to Figure 1.

Does the performance of Chebyshev prediction intervals depend on the type of data? To answer this question, further tests were made using subsets of the 111 time series. The results shown in Figure 3 are average percentages over all forecast horizons. There is little difference in Chebyshev performance between seasonal and nonseasonal data. However, there is an important difference in performance when the time series are classified by frequency of observation (annual, quarterly, and monthly). Chebyshev percentages are near targets for quarterly and monthly series but are substantially below targets in annual series. The problem is that most of the annual series in the Makridakis collection are too short to fit an adequate forecasting model. In the 20 annual series, the average number of fitted observations is 19.7. Nine of the annual series have 13 fitted observations and one has only 12. The quarterly and monthly series are much longer and average 38.4 and 68.4 fitted observations, respectively.

The sensitivity of Chebyshev prediction intervals to the number of fitted observations is shown in more detail by Figure 4. With less than 20 fitted observations, the Cheby-

FIGURE 2. Percentage of Post-Sample Observations Inside Prediction Intervals: 1,001 Time Series.

Forecast Leadtime	90% Target		95% Target	
	Chebyshev	Normal	Chebyshev	Normal
1	96.2%	80.1%	98.6%	86.2%
2	95.1	77.3	98.0	84.1
3	93.2	75.9	96.4	82.0
4	91.6	72.7	96.1	77.7
5	89.2	71.3	94.7	76.0
6	88.0	66.3	93.3	71.8
7	91.7	71.5	96.8	77.2
8	90.1	69.6	96.0	75.9
9	92.2	72.8	97.6	77.0
10	93.4	73.7	97.9	78.6
11	92.1	73.9	96.8	78.1
12	92.9	72.8	97.7	77.2
13	91.4	71.6	96.6	76.8
14	91.9	72.5	96.6	77.8
15	91.1	69.5	96.3	74.6
16	90.4	71.4	95.6	77.8
17	90.6	70.7	96.1	78.9
18	89.3	70.2	95.5	74.4
All	91.8%	72.7%	96.4%	78.2%

FIGURE 3. Performance by Type of Data.

Type of Data	Nbr. of Series	Nbr. of Forecasts	90% Target		95% Target	
			Chebyshev	Normal	Chebyshev	Normal
All data	111	1,528	90.8%	72.9%	95.8%	79.1%
Seasonal	60	990	91.7	72.2	96.7	77.8
Nonseasonal	51	538	89.0	73.1	93.1	79.2
Annual	20	120	78.3	58.3	83.3	63.3
Quarterly	23	184	88.0	59.2	95.7	71.7
Monthly	68	1,224	92.0	75.4	95.9	80.3

FIGURE 4. Performance by Number of Fitted Observations.

Nbr. of Fitted Obs.	Nbr. of Series	Nbr. of Forecasts	90% Target		95% Target	
			Chebyshev	Normal	Chebyshev	Normal
<20	17	110	71.8%	51.8%	82.7	57.3%
20–29	6	40	92.5	75.0	95.0	80.0
30–39	16	204	92.2	74.5	95.6	81.4
40–49	13	204	93.6	67.2	98.5	76.0
50–59	16	216	89.8	74.5	95.8	80.1
⩾60	43	754	91.9	74.9	95.4	79.8

shev intervals are not reliable. With 20 or more fitted observations, Chebyshev intervals are highly reliable.

There are at least two explanations for the good performance of the Chebyshev inequality. First, as discussed in Makridakis and Winkler (1985), the errors are not

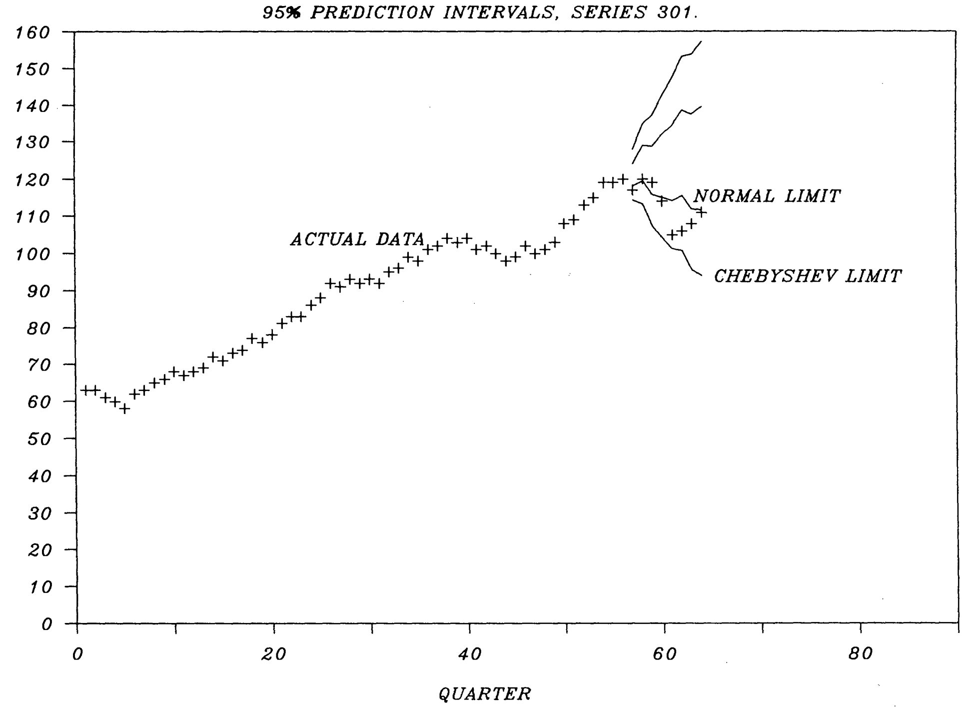

FIGURE 5

normally distributed in these data, regardless of the forecasting method. Second, in this research as well as in the M-competition, post-sample forecast errors were larger than within-sample errors. The Chebyshev inequality yields wider prediction intervals than the normal distribution in order to compensate for larger post-sample errors.

One reason that larger errors occur during the post-sample period is that time series frequently display changes in pattern or discontinuities. Except for Carbone and Makridakis (1986), this problem has been ignored in the literature. There are many series in the Makridakis data that display changes in the direction of trend after the forecasts are made. Figure 5 illustrates this problem for the quarterly time series (number 301) that was analyzed in Carbone and Makridakis (1986). This series is composed of 64 observations, with a weak seasonal pattern and an unstable trend. An exponential smoothing model with damped trend and multiplicative seasonality was fitted to the series, using quarters 1–56. During model-fitting, prediction intervals were developed for quarters 57–64. The Chebyshev intervals accommodate the drastic change in trend whereas the normal intervals do not. All eight post-sample observations are inside the Chebyshev intervals, while only three are inside the normal intervals.

4. Conclusions

The normal distribution is a standard assumption for determining prediction intervals in time series forecasting. However, normal prediction intervals are dangerously misleading. Normal intervals are too small regardless of the forecast leadtime, the type of data, or the number of fitted observations. This finding is consistent with other work using a variety of different forecasting methods by Williams and Goodman (1971), Lusk and Belhadjali (1986), and Makridakis and Hibon (1986).

Since it appears to be impossible to find any adequate distribution for forecast errors, the Chebyshev inequality seems an obvious alternative for setting bounds on the errors.

In the Makridakis data, the Chebyshev yields accurate prediction intervals. Compared to the normal distribution, the primary reason that the Chebyshev works is that it increases prediction intervals to compensate for larger errors during the post-sample period. There is no guarantee that this strategy will perform as well in other data. However, deteriorating accuracy during the post-sample period has been observed in numerous empirical studies and is by now a generally accepted outcome in time series forecasting.

In order to apply the Chebyshev inequality, variances of the errors by forecast lead-time are necessary. The method of estimating variances in this paper is simple and does not rely on any assumptions about the validity of the model or the form of the generating function for the time series. The method is certainly time-consuming but it is less so than the Williams and Goodman approach.[1]

[1] The author is indebted to two anonymous referees and Professor Lynn Lamotte of the University of Houston for helpful suggestions on this research.

References

BOX, G. AND G. JENKINS, *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, San Francisco, 1976.

BROWN, R., *Smoothing, Forecasting, and Prediction of Discrete Time Series*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

CARBONE, R. AND S. MAKRIDAKIS, "Forecasting When Pattern Changes Occur Beyond the Original Data," *Management Sci.*, 32 (March 1986), 257–271.

GARDNER, E. S., *Autocast Exponential Smoothing System User's Guide*, Core Analytic, Bridgewater, NJ, 1986.

——— AND E. MCKENZIE, "Forecasting Trends in Time Series," *Management Sci.*, 31 (October 1985), 1237–1246.

LUSK, E. AND K. BELHADJALI, "Prediction Intervals for ARIMA Models for the Makridakis 111 Series," *Internat. J. Forecasting*, forthcoming (1986).

MAKRIDAKIS, S. ET AL., "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *J. Forecasting*, 2 (April–June 1982), 111–153.

——— AND M. HIBON, "Prediction Intervals: an Empirical Investigation for the Series in the *M*-Competition," *Internat. J. Forecasting*, forthcoming (1986).

——— AND R. WINKLER, "The Sampling Distribution of Post-sample Forecast Errors," INSEAD working paper, 1985.

MCKENZIE, E., "A Comparison of Some Standard Seasonal Forecasting Systems," *The Statistician*, 25 (January 1976), 3–14.

———, "General Exponential Smoothing and the Equivalent ARMA Process," *J. Forecasting*, 3 (July–September 1984), 333–344.

———, "Error Analysis for Winters' Additive Seasonal Forecasting System," *Internat. J. Forecasting*, 2 (April–June 1986), 373–382.

SWEET, A., "Computing the Variance of the Forecasting Error for the Holt-Winters Seasonal Models," *J. Forecasting*, 4 (October–December 1985), 235–243.

WILKS, S. S., *Mathematical Statistics*, Wiley, New York, 1962.

WILLIAMS, W. AND M. GOODMAN, "A Simple Method of Constructing Confidence Limits for Economic Forecasts," *J. Amer. Statist. Assoc.*, 66 (December 1971), 752–754.