Narrowness Of Prediction Intervals Revisited

Bernard J. Morzuch and P. Geoffrey Allen
Department of Resource Economics
University of Massachusetts
Amherst, MA 01003 USA
(tel: (413) 545-5718)
(fax: (413) 545-5853)
morzuch@resecon.umass.edu
allen@resecon.umass.edu

Presented at the 20th Annual Symposium on Forecasting
Lisbon, Portugal, June 21-24, 2000

Narrowness Of Prediction Intervals Revisited

Abstract

Prediction intervals are frequently too narrow. This problem, however, receives minimal attention in the forecasting literature. Probabilistic forecasts have a number of advantages over more common point forecasts, most notably the additional information conveyed. As forecasters we ought to pay more attention to their historically poor calibration in post-sample prediction periods.

In the past, we have developed and presented strategies for promoting well-calibrated prediction intervals under a variety of model specifications and using different types of data series. These strategies employed parametric procedures. Their success has been limited at best but still better than other studies dealing with this problem.

We attempt to improve our previous work by incorporating bootstrapping methods into our calibration procedures. Care in model specification, diagnostic testing, and judicious use of resampling procedures for generating key statistics needed for prediction intervals may promote better calibrated prediction intervals.

Narrowness Of Prediction Intervals Revisited

Introduction

It is well known that prediction intervals (PIs) constructed around point forecasts are in general too narrow. Chatfield (1993) provides a comprehensive review of the various methods of calculating prediction intervals and discusses (p. 131) seven reasons for this problem occurring. The two most likely reasons are: (1) the wrong model has been identified; that is, the data generating process is not well-approximated by the forecasting model selected; and (2) the underlying model changes in the forecast period; that is, based upon some model specification, a series may appear well-behaved within-sample but takes a departure from specification in the ex post period.

Regarding the first issue above, series that are not well-behaved within sample are frequently encountered. (Then it should be no surprise when post-sample prediction intervals are too narrow!) Yet, appropriate specification tests are not widely performed. Regarding the second issue, even if a model passes specification tests and the series appears well-behaved within-sample, a specification tends to break down as the forecast horizon lengthens. Here, there is little the modeler can do.

Makridakis et al. (1987) instigated our interest in this area by measuring the percentage of actual values that fall outside the prediction intervals constructed for each of 15 methods used in the M-Competition (Makridakis, et al., 1982). Each of the 15 methods was applied to all 1001 time series of the M-competition. The make-up of the 1001 series was as follows: 181 yearly, 203 quarterly, and 617 monthly.

The percentage of observations for all series outside the prediction intervals for each method's within-sample period was close to that postulated theoretically; e.g., a 95% PI resulted

in about 5% of the observations to be outside the constructed interval. However, in the great majority of cases the percentage of observations outside the prediction interval for each method's post-sample period was much larger than that postulated theoretically. The percentage became even larger as the forecasting horizon became longer.

Their conclusions were as follows:

(1) They suggested that these larger-than-expected percentages may be due to changes in the pattern of the data, systematic biases in the forecasting method used, lack of normality, heteroscedasticity, and serial correlation.

(2) They recommended the need to construct realistic, non-symmetrical prediction intervals.

(3) Ways must be found to deal with series whose established pattern may change during the ex post period.

Literature

Researchers responded to these conclusions in various ways. Regarding (1), Makridakis and Winkler (1989) made detailed comparisons of forecast errors for 111 series of the M-competition separately for the within-sample (model fitting) period and ex post (average of 1-6 steps ahead) period. They used eight methods and calculated mean percentage error, skewness, kurtosis, and autocorrelations of lags 1 through 4 of the fitted errors for the within-sample period. The authors did not attempt any calibration measurements; we do not know whether or not any or all of the methods gave forecast PIs that were too narrow.

Regarding (2), Chatfield (1993, p. 124) points out that the one-step-ahead forecast (and by extension the h-steps-ahead forecast) will not in general be normal, even for a linear model with normally distributed innovations (Phillips, 1979). The empirical method of Williams and Goodman (1971) and the *ad hoc* method of Gardner (1988) "work" in the sense of bringing the

prediction intervals, on average, in line with the distribution of actual observations. Without these adjustments, prediction intervals calculated by true-model or approximation formulas are too narrow, regardless of forecast method, data frequency, or length of forecast horizon.

Regarding (3), it is impossible to gaze into the future and discern a future pattern change in a series, irrespective of any model that is used or any judgement that is exercised.

Because of lack of success in ex post forecasting, the current and conventional wisdom appears to be that all forecasting methods perform equally badly. The corollary is not even to bother with within-sample testing. In fact, because of the persistent lack of success in these areas, Fildes and Makridakis (1995) emphasized the need for a theoretical framework in which to develop improved forecasting methods and establish effective selection criteria.

Previous Work By Allen and Morzuch

Allen and Morzuch (1996) responded to the challenge by emphasizing the importance of within-sample specification tests on a model before making post-sample predictions. We used the NAIVE2 method (no change forecast on deseasonalized data) on all 1001 series in the Makridakis competition (181 yearly data series, 203 quarterly data series, and 617 monthly data series) to show how results from within-sample specification tests determine the adequacy of post-sample prediction intervals. Attention focused on testing the residual of the random walk model for zero mean, normality, homoscedasticity, and autocorrelation.

Based upon within-sample estimation results using this method, we divided the population of time series into two groups: the first group consisted of those series on which a misspecified model had been applied; the second group consisted of those series for which we could not show that the model was misspecified. The question was whether models that were known to be misspecified had worse prediction intervals than models that were not misspecified.

(Worse meant that ex post forecast errors were not calibrated uniformly in the deciles of the normal distribution.)

Several conclusions were drawn from this exercise. First, using a forecasting model only when it was appropriate did improve post-sample performance. Specifically, using the naive method only on series that passed all within-sample tests gave better results than using it on series that failed one or more tests. While this improvement held only for a maximum of two, seven, and six forecast horizons out of six, eight, and eighteen horizons for yearly, quarterly, and monthly series, respectively, it showed that discrimination based on within-sample tests can be expected to improve forecasting performance.

At the time, we did not perform any parameter constancy tests to address issues relating to trends in series. One of the glaring violations was some type of trending in many of the series, for which the random walk model was obviously deficient. Thus, our next endeavor (Allen and Morzuch, 1997) focused on this problem. We introduced an additional layer of realism by considering and testing parameter constancy in the context of the random walk plus drift model. This, combined with the battery of tests used in the 1996 study, became our benchmark for classifying series. Thus, if a series was generated by a random walk plus drift process, results from our within-sample specification tests could be used to determine adequate post-sample prediction intervals.

As in the previous study, using a forecasting model – in this case the random walk plus drift – only when it was appropriate resulted in improved post-sample performance. However, results were not as striking as we had anticipated in terms of calibration of the prediction intervals. In the 1997 study, we presented a variety of plausible explanations for, in our opinion, these less-than-favorable results. Indeed, an additional parameter to estimate when moving to the

random walk plus drift model translates into more uncertainty. This resulted in the potential for increased forecast error and is reflected in a larger standard deviation of the forecast error itself.

While our results in both studies suggested that we may be heading in the right direction, one consideration that has plagued us and every other researcher when it comes to calibration relates to obtaining reliable estimates of the standard deviation of the forecast error. Thus, we now refocus on this basic element of the prediction interval itself.

Resampling Procedures

We keep things simple by returning to the random walk model. We begin with the premise that the standard deviation of the forecast error for h-steps ahead is the within-sample standard deviation of the error ("s") multiplied by $\sqrt{h}$ (Chatfield, 1993, p. 124). Thus, for example, the one-step ahead standard deviation of the forecast error is $s \cdot \sqrt{h}$ $s \cdot \sqrt{1}$ s. Also, recall that the sample variance is an unbiased estimator for σ^2, but s is a biased estimator for σ.

All of our previous analyses required an extensive battery of statistical tests (presented again in the next section), for which we made common parametric assumptions throughout. The data series used in our analyses had both large and small numbers of observations.

Two general caveats exist when constructing confidence and prediction intervals. First, they will be exact only when we assume that the population from which we sample is normal. Without the normality assumption, results are large sample results based upon the approximate normality of the sample variance or the sample standard deviation. The second caveat relates to the bias of the sample standard deviation in small samples. (The first caveat implies a statement about the second).

Resampling procedures, of which the bootstrap is a candidate, present an avenue for dealing with the two caveats presented above. Most often recognized as a method for deriving a

summary statistic when no analytic formula exists, its method is such that it avoids distributional assumptions entirely, and it provides an outlet for dealing with estimation bias in small samples. These procedures have become quite attractive with the advent of high speed computers. Good background references are Diaconis and Efron (1983), Efron and Tibshirani (1993), Davison and Hinkley (1997), and Chernick (1999).

Throughout our analyses, we had the choice of how and to what we might apply the bootstrap. As indicated above, we developed and analyzed prediction intervals based upon an extensive battery of parametric procedures. In our present analysis, each procedure could have been amended with a resampling counterpart, resulting in a complete nonparametric approach.

We chose not to take this direction but to build toward it incrementally, starting with a nonparametric way of (hopefully) improving the estimate of the standard deviation while keeping our parametric assumptions and procedures in tact for the remainder of the analyses. This blend of parametric and nonparametric techniques is sometimes referred to as the semiparametric (or semi nonparametric) approach. One reason we chose to proceed incrementally along these lines is that a direct comparison could then be made with our previous work (Allen and Morzuch, 1996) by changing only one item and letting everything else remain as it was.

As the bootstrap procedure relates to our particular situation, it was applied straightforwardly as follows. Effectively, for each within-sample time series, we sampled the empirical distribution of the variable of interest, in our case the error term or the first difference. Each data series was of different length, resulting in a different number of first differences. Suppose that, for any series, there are T such first differences in general. We proceed as follows. Start with the first series, select an error term, record its value, replace it, and repeat this process T-1 additional times. For all T observations, calculate a mean and a standard deviation. Repeat

this entire procedure B times in total, and concentrate on the empirical distribution of the B standard deviations that are calculated. (The reason for concentrating on the standard deviation is because it is the statistic of interest for prediction intervals and post-sample calibration. Now, pay attention to the *mean* of the B standard deviations. This will be our estimate of the true standard deviation. Finally, repeat this process for each remaining series.

Notice that this procedure does not depend upon any parametric assumptions. It tends to perform better as more bootstrap samples are taken, i.e., the larger is B.

Within-Sample Tests

Following the same procedure as Makridakis, et al. (1987), we subjected each of the 181 yearly data series, the 203 quarterly data series, and 617 monthly data series to the NAIVE2 method. In using the NAIVE2 method, we assume that the random walk model represents the data generating process. Once we assume that this is the correct model, we are first obligated to test the assumptions within sample. Using a monthly data series as an example, we proceed as follows.

(1) Deseasonalize the data.

(2) Take first differences.

(3) Calculate the mean of (2), the standard error of the mean, and the t-value. This becomes a test of the null hypothesis that the mean of the within-sample errors is zero.

(4) Calculate measures of skewness and kurtosis of the error according to the suggestions of D'Agostino, et al. (1990). Test the null hypothesis of normality of the error versus the alternative of nonnormality due to skewness and then the null hypothesis of normality of the error versus nonnormality due to nonnormal kurtosis.

(5) Combine the skewness and kurtosis measures calculated in (4) to produce an omnibus test

of normality. This test is able to detect deviations from normality due to either skewness or kurtosis. The test statistic itself has approximately a chi-squared distribution with two degrees of freedom when the errors are normally distributed (D'Agostino, et al., 1990).

(6) Calculate the Ljung-Box Q-statistic for up to 12 lags, if the data permit, to test for autocorrelation of the error term.

(7) Use an ARCH model to test for (dynamic) heteroscedasticity. We regress each squared residual on a constant and on the squares of four lagged values of itself. (The number of lags is purely arbitrary). If there are no ARCH effects, the estimated coefficients on the squared residuals should be zero. Thus, the regression will have little explanatory power, and the coefficient of determination will be low. Under the null hypothesis of no heteroscedasticity (no ARCH effects) for a sample of T residuals, the test statistic is TR^2. This test statistic converges to χ^2_K where K is the number of lags. In our case, K=4. It should be noted that this is a Lagrange Multiplier test.

(8) For each test statistic in (3) through (7), calculate its p-value.

Next, we pay attention to whether or not the series passes all of the diagnostic tests. Each test was conducted using the 5% level of significance. If the series passes all tests, we do not have evidence to suggest that the data generating process is different from a random walk. If it fails even one test, we have evidence to suggest that the data generating process is different from a random walk. Importantly, the former result suggests that the random walk model may be appropriate for ex post forecasting while the latter result cautions us not to use the random walk for ex post forecasting.

We repeat this process for the remaining 616 monthly data series and form two groups: one consisting of those series that pass all within-sample tests and another consisting of those

series that fail one or more within-sample tests. We repeat this entire procedure for all annual series and for all quarterly series. Table 1 lists the six groupings that result from within-sample testing.

Post-Sample Calibration

We can now answer the question: Does the battery of diagnostic tests on the within-sample model provide any useful information for ex post forecasting? The test is performed separately by frequency of data, by grouping, and by lead time. For example, consider the behavior of the one-step-ahead forecast error of a member of the monthly data series in the group that has passed all within-sample tests. The forecast error is a random variable and if it is normally distributed around a mean of zero or, for test purposes, uniformly distributed within the deciles of the normal distribution, our errors are properly calibrated. Neither too few nor too many errors are in the tails, suggesting that we may have discovered the data generating process for these series.

Continuing with the example, we developed the calibration test as follows.

(1) In our 1996 and 1997 studies, we applied the random walk model to the series and calculated the within-sample standard deviation of the error. In the present study, we bootstrap the within-sample standard deviation of the error. The bootstrap estimate was based on 250 iterations.

(2) The standard deviation of the forecast error for h-steps ahead is the item calculated in (1) times $\sqrt{h}$. (Chatfield, 1993, p. 124). For one-step ahead, $\sqrt{h}$ 1.

(3) Since within-sample tests did not permit us to reject both the null hypothesis of zero mean for the error and the null hypothesis of normality, we can use the estimate in (2) to construct deciles (and decile boundaries) under the normal curve.

(4) Perform the one-step ahead forecast, calculate the one-step ahead forecast error, and record its position within the appropriate decile boundaries using (3) above.

(5) Repeat (1)-(4) for each of the remaining series in the group.

(6) Tally the results over all of the series.

(7) If our within-sample model is useful for post-sample forecasting, we would expect a uniform distribution of forecast errors in the deciles. Perform a χ^2 goodness-of-fit test for the tallied results.

(8) Repeat (1) - (7) for each forecasting horizon.

(9) Repeat (1) - (8) for quarterly and annual data series for those groups that passed all within-sample tests.

Results

Table 2 provides a detailed summary of the results of the within-sample diagnostic tests. The tests would normally be performed on the residual errors from model fitting (or the one-step-ahead forecast errors, within-sample). With the random walk, the differenced series is equivalent. For each data type, the table shows the number of series that pass all tests (column 2) and the number that pass all tests but zero mean (column 3), which indicates a random walk plus drift. The remaining columns present frequencies of failure for each specific test.

Monthly and quarterly series either had to be deseasonalized using the factors provided by Makridakis, et al. (1982) or they were nonseasonal from the start; that is, they had a seasonal factor of one. We report the results only for nonseasonal series since estimation of smoothing parameters should be conducted in the context of within-sample diagnostic testing.

Table 3 provides post-sample results for the annual data based on the within-sample tests. It shows the percentage of forecast errors falling in each decile of the distribution for series that

pass all within-sample tests. Note also that results are based upon bootstrapped estimates for the post sample standard deviation.

A chi-squared goodness of fit test was performed to determine if the ex post forecasts were distributed uniformly in the deciles for each of the six forecast horizons. For a forecast horizon of one period, uniformity could not be rejected ($p = 0.57$). For two periods, uniformity could not be rejected at the 0.10 level; i.e., $p = 0.10$. However, evidence was strong to suggest rejection for leads of three through six periods. Thus, unlike previous studies which lump outcomes of within-sample tests together, our discriminate procedure suggests that cautious within-sample testing may result in properly calibrated ex post forecasts for up to two forecast horizons.

Table 5 provides post-sample results for the nonseasonal quarterly data based upon within-sample tests. The quarterly nonseasonal group of series that passed all within-sample tests gave too small a sample size for a reliable goodness-of-fit test. Creating pentiles (five groups) out of our deciles does give a sufficiently large sample size to conduct the tests. Test results indicated that uniformity of the forecast error distribution could not be rejected for seven of the eight forecast horizons.

Table 7 provides post-sample results based upon within-sample tests using monthly nonseasonal data. Uniformity of the forecast error distribution could not be rejected for six of the 18 forecast horizons when all within-sample tests were passed. One curious result appeared in the distribution of those series that passed the within-sample tests. Larger frequencies tended to be located in the left tail. One major reason that uniformity was rejected for 12 of the 18 forecast horizons was excessive counts occurring in the left tail of the distribution.

Did the bootstrap procedure lead to improvement relative to the 1996 study?

Unfortunately for us, the answer is no. Comparing Table 3 results with the 1996 study where no bootstrapping was performed (Table 4), we notice that the decile distributions of forecast errors are virtually identical between tables. The same holds true when making comparisons for the quarterly series (Tables 5 and 6) and for the monthly series (Tables 6 and 7).

Discussion

Was the bootstrapping procedure a waste of time? In terms of the virtual identicalness between respective pairs of tables, the answer appears to be yes. At the outset of this endeavor, however, we never would have thought this to be the case. From the perspective of information conveyed when using this procedure, we regard the experience as invaluable. The sources previously cited about bootstrapping paint a fairly clear picture about its benefits and payoffs. Yet we saw none, and this is counter to the conventional wisdom. Indeed, one reason that we believed that we would see a difference in results was because the bootstrapped standard deviations were different than their non-bootstrapped counterparts for a reasonable percentage of the series. Two issues follow from this result.

We did not take the time to compare the change in standard deviation relative to series length. Discrimination between short and long series can have significant consequences. Secondly, we did not address the important issues of bias correction and acceleration as they relate to each individual series. Indeed, dealing with these issues for each individual series is another layer of significant time consumption, but it appears necessary.

References

Allen, P.G. and B.J. Morzuch, 1996. "So Why Are Prediction Intervals (Almost) Always Too Narrow?" Presented at the 16th Annual Symposium on Forecasting, Istanbul Turkey, June 24-26, 1996.

Allen, P.G. and B.J. Morzuch, 1997. "Further Evidence On Why Prediction Intervals Are (Almost) Always Too Narrow." Presented at the 17th Annual Symposium on Forecasting, Barbados, June 19-21, 1997.

Chatfield, C., 1993. "Calculating Interval Forecasts." *Journal of Business and Economic Statistics*, 11, 121-135.

Chernick, M.R., 1999. Bootstrap Methods: A Practitioner's Guide. New York: Wiley.

D'Agostino, R.B., A. Belanger and R.D D'Agostino, Jr., 1990. "A Suggestion Of Using Powerful And Informative Tests Of Normality." *American Statistician*, 44, 316-321.

Davison, A.C. and D.V. Hinkley, 1997. Bootstrap Methods And Their Application. Cambridge, U.K.; New York: Cambridge University Press.

Diaconis, P. and B. Efron, 1983. "Computer-Intensive Methods In Statistics." Scientific American, 248, 116-130.

Efron, B. and R.J. Tibshirani, 1993. An Introduction To The Bootstrap. New York: Chapman & Hall.

Fildes, R. and S. Makridakis, 1995. "The Impact Of Empirical Accuracy Studies On Time Series Analysis And Forecasting." *International Statistical Review*, 63, 289-308.

Gardner, E.S., 1998. "A Simple Method Of Computing Prediction Intervals For Time Series Forecasts." *Management Science*, 34, 541-546.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, 1982. "The Accuracy Of Extrapolation (Time Series) Methods: Results Of A Forecasting Competition." *Journal of Forecasting*, 1, 111-153.

Makridakis, S., M. Hibon, E. Lusk, and M. Belhadjali, 1987. "Confidence Intervals: An Empirical Investigation Of The Series Of The M-Competition." *International Journal of Forecasting*, 3, 489-508.

Makridakis, S. and R. L. Winkler, 1989. "Sample Distributions Of Post-Sample Forecasting Errors." *Applied Statistician*, 38, 331-342.

Phillips, P.C.B., 1979. "The Sampling Distribution Of Forecasts From A First-Order Autoregression." *Journal of Econometrics*, 9, 241-261.

Williams, W.H. and 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.

Table 1: Groupings Of Series Based On Within-Sample Diagnostic Tests.

Group	Data Series	Test Outcome
1	Annual	Pass
2	Annual	Fail
3	Quarterly	Pass
4	Quarterly	Fail
5	Monthly	Pass
6	Monthly	Fail

Table 2: Number Of Series Passing All Within-Sample Diagnostic Tests, Or All But Zero Mean, And Number Failing Each Test For Different Data Types, Using First Differences Of The Original Data.

		Pass		Fail zero mean	Fail Normality			Fail auto	Fail heter
Data type	Total number	all	all but zero mean		skew	kurt	omni		
Annual	181	52	56	91	35	37	37	30	13
Quarterly nonseasonal	114	19	16	55	35	37	40	30	13
Monthly nonseasonal	209	38	8	37	71	99	108	96	62

Tests (and distribution under the null hypothesis)
zero mean: sample mean equals zero (t)
skew: D'Agostino skewness measure based on third moment about mean (z)
omni: D'Agostino K^2 statistic based on third and fourth moments about the mean (χ^2_2)
auto: Ljung-Box Q statistic for autocorrelation (χ^2_{12})
heter: Engle's ARCH test for dynamic heteroscedasticity on 4 lags of squared differenced observations (χ^2_4)

Table 2 is taken from Allen and Morzuch (1996). Quarterly deseasonalized and monthly deseasonalized results are excluded.

Table 3: Annual Series (n=52), Decile Distribution Of Forecast Errors (In Percents)
Using **Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	10	4	11	13	10	6	15	4	13	13
2	6	4	11	8	10	6	10	10	11	25
3	10	8	2	11	8	4	6	0	17	34
4	12	8	4	6	8	6	6	11	8	32
5	10	6	6	6	8	8	11	2	8	36
6	10	4	6	6	2	13	6	11	4	38

Table 4: Annual Series (n=52), Decile Distribution Of Forecast Errors (In Percents)
Using **Non-Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	10	4	10	15	10	6	15	6	13	12
2	6	4	12	8	10	6	12	8	13	23
3	8	10	2	12	8	4	6	2	15	35
4	10	10	2	8	8	6	6	12	10	31
5	10	6	4	6	8	8	12	4	8	35
6	10	2	6	8	2	13	6	13	2	38

Source for Table 4 is Allen and Morzuch (1996)

Table 5: Quarterly Nonseasonal Series (n=19), Decile Distribution Of Forecast Errors (In Percents) Using **Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	11	11	16	16	5	5	0	11	16	11
2	16	10	5	5	16	26	0	0	11	11
3	11	0	21	11	11	5	11	11	11	11
4	11	5	11	5	5	32	5	11	5	11
5	11	16	0	0	5	16	16	21	0	16
6	21	0	5	5	16	16	11	5	11	11
7	11	11	5	16	11	5	11	11	11	11
8	5	16	5	16	5	21	5	11	0	16

Table 6: Quarterly Nonseasonal Series (n=19), Decile Distribution Of Forecast Errors (In Percents) Using **Non-Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	11	11	16	16	5	5	0	16	11	11
2	16	11	5	0	21	26	0	0	11	11
3	11	0	16	16	11	5	11	11	11	11
4	11	5	11	5	5	32	5	11	5	11
5	5	21	0	0	5	16	16	21	0	16
6	21	0	5	5	16	16	11	5	11	11
7	5	16	5	16	11	11	5	11	11	11
8	5	11	11	16	5	21	5	11	0	16

Source: Allen and Morzuch (1996)

Table 7: Nonseasonal Monthly Series (n=38), Decile Distribution Of Forecast Errors (In Percents) Using **Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	23	16	8	11	11	8	3	8	3	11
2	16	16	11	16	5	8	0	11	5	13
3	16	21	16	11	5	18	3	3	5	3
4	21	26	11	0	5	11	5	8	11	3
5	32	16	16	0	3	8	8	5	11	3
6	29	21	11	3	3	8	11	5	5	5
7	16	18	24	8	0	13	5	11	3	3
8	24	21	5	16	8	8	3	5	5	5
9	26	16	11	8	11	8	8	5	5	3
10	37	16	0	8	5	11	5	11	3	5
11	37	13	5	5	11	8	8	8	3	3
12	29	16	8	3	11	18	5	8	0	3
13	18	23	8	5	13	11	8	11	0	3
14	16	23	13	8	13	11	8	3	3	3
15	16	18	11	13	18	8	8	5	0	3
16	8	21	11	16	16	11	11	5	0	3
17	8	21	18	16	11	11	8	5	0	3
18	13	18	18	16	8	11	11	0	3	3

Table 8: Nonseasonal Monthly Series (n=38), Decile Distribution Of Forecast Errors (In Percents) Using **Non-Bootstrapped** Standard Deviation Of Forecast Error In Ex Post Period

Steps ahead	Deciles									
	0 to .1	.1 to .2	.2 to .3	.3 to .4	.4 to .5	.5 to .6	.6 to .7	.7 to .8	.8 to .9	.9 to 1
1	24	13	11	11	11	11	3	8	3	11
2	16	16	11	16	5	5	3	8	8	11
3	16	21	16	11	5	5	3	3	5	3
4	18	29	11	0	5	5	3	8	11	3
5	32	16	16	0	3	3	5	5	11	3
6	24	26	11	3	3	3	11	5	5	5
7	16	18	24	5	0	3	5	11	3	3
8	24	21	5	16	8	8	5	3	5	5
9	26	16	11	8	11	11	8	5	5	3
10	34	18	0	8	5	5	5	11	3	5
11	37	13	5	5	11	11	8	8	3	3
12	29	16	8	3	11	11	5	8	0	3
13	18	24	8	5	13	13	8	8	0	3
14	16	24	13	5	13	16	8	3	3	3
15	16	18	11	13	18	18	8	5	0	3
16	8	21	11	16	16	16	11	5	0	3
17	5	24	18	16	11	11	8	5	0	3
18	11	21	16	18	8	8	11	0	3	3

Source: Allen and Morzuch (1996)