Measuring forecast accuracy

Rob J Hyndman

Monday, 31 March 2014

Everyone wants to know how accurate their forecasts are. Does your forecasting method give good forecasts? Are they better than the competitor methods?

There are many ways of measuring the accuracy of forecasts, and the answers to these questions depends on what is being forecast, what accuracy measure is used, and what data set is used for computing the accuracy measure. In this chapter, I will summarize the most important and useful approaches.

1 Training and test sets

It is important to evaluate forecast accuracy using genuine forecasts. That is, it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when estimating the model. When choosing models, it is common to use a portion of the available data for testing, and use the rest of the data for estimating (or "training") the model. Then the testing data can be used to measure how well the model is likely to forecast on new data.

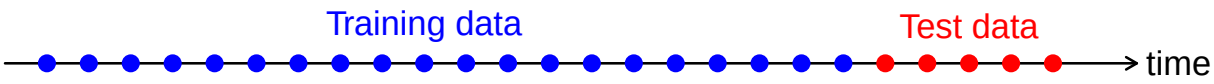

Figure 1: *A time series is often divided into training data (used to estimate the model) and test data (used to evaluate the forecasts).*

The size of the test data set is typically about 20% of the total sample, although this value depends on how long the sample is and how far ahead you want to forecast. The size of the test set should ideally be at least as large as the maximum forecast horizon required.

[1] This chapter is based on Section 2.5 of *Forecasting: principles and practice* by Rob J Hyndman and George Athanasopoulos, available online at www.otexts.org/fpp/2/5, and used with permission.

The following points should be noted.

- A model which fits the data well does not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is as bad as failing to identify the systematic pattern in the data.

Some references describe the test data as the "hold-out set" because these data are "held out" of the data used for fitting. Other references call the training data the "in-sample data" and the test data the "out-of-sample data".

2 Forecast accuracy measures

Suppose our data set is denoted by $y_1,\dots,y_T$, and we split it into two sections: the training data $(y_1,\dots,y_N)$ and the test data $(y_{N+1},\dots,y_T)$. To check the accuracy of our forecasting method, we will estimate the parameters using the training data, and forecast the next $T-N$ observations. These forecasts can then be compared to the test data.

The h-step-ahead forecast can be written as $\hat{y}_{N+h|N}$. The "hat" notation indicates that it is an estimate rather than an observed value, and the subscript indicates that we are estimating y_{N+h} using all the data observed up to and including time N.

The forecast errors are the difference between the actual values in the test set and the forecasts produced using only the data in the training set. Thus

$$e_t = y_t - \hat{y}_{t|N}, \qquad \text{for } t = N+1,\dots,T.$$

Scale-dependent errors

These errors are on the same scale as the data. For example, if y_t is sales volume in kilograms, then e_t is also in kilograms. Accuracy measures that are based directly on e_t are therefore scale-dependent and cannot be used to make comparisons between series that are on different scales.

The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

$$\text{Mean absolute error: MAE} = \text{mean}(|e_i|),$$

$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_i^2)}.$$

When comparing forecast methods on a single data set, the MAE is popular as it is easy to understand and compute.

Percentage errors

The percentage error is given by $p_t = 100e_t/y_t$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance between different data sets. The most commonly used measure is:

$$\text{Mean absolute percentage error: MAPE} = \text{mean}(|p_t|).$$

Measures based on percentage errors have the disadvantage of being infinite or undefined if $y_t = 0$ for any observation in the test set, and having extreme values when any y_t is close to zero.

Another problem with percentage errors that is often overlooked is that they assume a scale based on quantity. If y_t is measured in dollars, or kilograms, or some other quantity, percentages make sense. On the other hand, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales, because these are not measuring a quantity. One way to think about it is that percentage errors only make sense if changing the scale does not change the percentage. Changing y_t from kilograms to pounds will give the same percentages, but changing y_t from Fahrenheit to Celsius will give different percentages.

Scaled errors

Scaled errors were proposed by Hyndman and Koehler (2006) as an alternative to using percentage errors when comparing forecast accuracy across series on different scales. A scaled error is given by $q_t = e_t/Q$ where Q is a scaling statistic computed on the training data. For a non-seasonal time series, a useful way to define the scaling statistic is the mean absolute difference between consecutive observations:

$$Q = \frac{1}{N-1}\sum_{j=2}^{N}|y_j - y_{j-1}|.$$

That is, Q is the MAE for naïve forecasts computed on the training data. Because the numerator and denominator both involve values on the scale of the original data, q_t is independent of the

scale of the data. A scaled error is less than one if it arises from a better forecast than the average naïve forecast computed on the training data. Conversely, it is greater than one if the forecast is worse than the average naïve forecast computed on the training data. For seasonal time series, a scaling statistic can be defined using seasonal naïve forecasts:

$$Q = \frac{1}{N-m}\sum_{j=m+1}^{N} |y_j - y_{j-m}|.$$

The *mean absolute scaled error* is simply

$$\text{MASE} = \text{mean}(|q_j|) = \text{MAE}/Q.$$

The value of Q is calculated using the training data because it is important to get a stable measure of the scale of the data. The training set is usually much larger than the test set, and so allows a better estimate of Q.

Example: Australian quarterly beer production

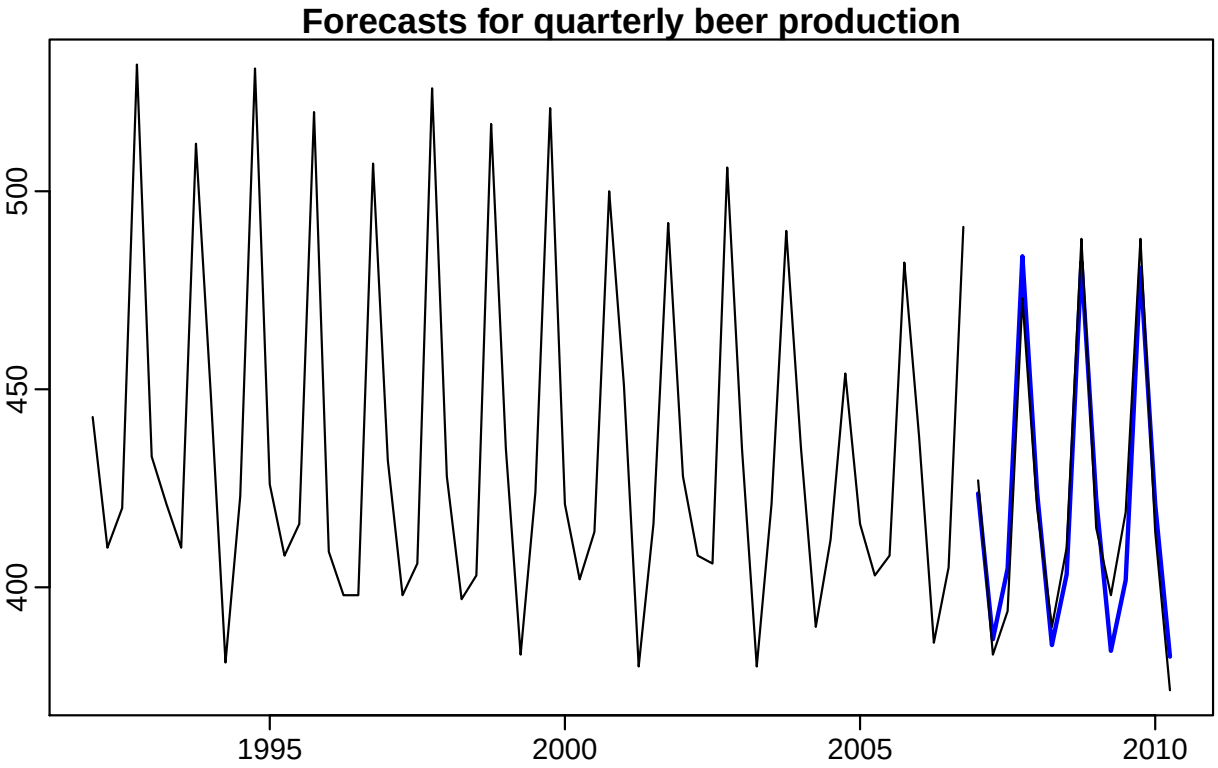

Figure 2: *Forecasts of Australian quarterly beer production using an ARIMA model applied to data up to the end of 2005. The black line shows actual values (in the training and test data sets) while the blue line shows the forecasts.*

	Actual	Forecast	Error	Percent. error
2007 Q1	427	423.69	3.31	0.78
2007 Q2	383	386.88	−3.88	−1.01
2007 Q3	394	404.71	−10.71	−2.72
2007 Q4	473	483.59	−10.59	−2.24
2008 Q1	420	423.81	−3.81	−0.91
2008 Q2	390	385.42	4.58	1.17
2008 Q3	410	403.25	6.75	1.65
2008 Q4	488	482.13	5.87	1.20
2009 Q1	415	422.35	−7.35	−1.77
2009 Q2	398	383.96	14.04	3.53
2009 Q3	419	401.79	17.21	4.11
2009 Q4	488	480.67	7.33	1.50
2010 Q1	414	420.89	−6.89	−1.66
2010 Q2	374	382.50	−8.50	−2.27
MAE			7.92	
RMSE			8.82	
MAPE				1.89%
MASE			0.54	

Table 1: *Accuracy measures computed from ARIMA forecasts for the 14 observations in the test data.*

Figure 2 shows forecasts for quarterly Australian beer production[1] An ARIMA model was estimated on the training data (data from 1992 to 2006), and forecasts for the next 14 quarters were produced. The actual values for the period 2007–2010 are also shown.

The forecast accuracy measures are computed in Table 1. The scaling constant for the MASE statistic was $Q = 14.55$ (based on the training data 1992–2006).

3 Time series cross-validation

For short time series, we do not not want limit the available data by splitting some off in a test set. Also, if the test set is small, the conclusions we draw from the forecast accuracy measures may not be very reliable. One solution to these problems is to use "time series cross-validation".

In this approach, we use many different training sets, each one containing one more observation than the previous one. Figure 3 shows the series of training sets (in blue) and test sets (in red). The forecast accuracy measures are calculated on each test set and the results are averaged across all test sets (adjusting for their different sizes).

[1]The data were obtained from the Australian Bureau of Statistics, Cat.No.8301.0.55.001.

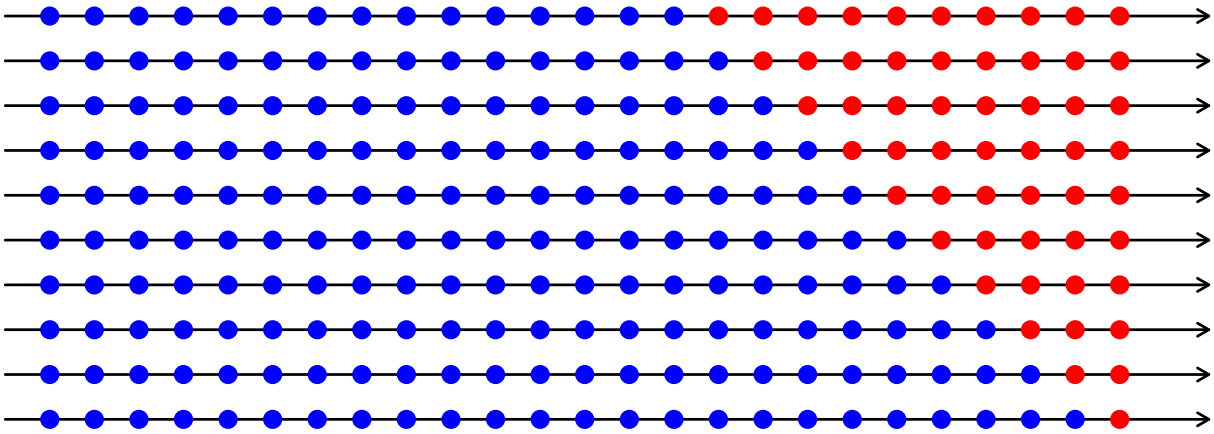

Figure 3: *In time series cross-validation, a series of training and test sets are used. Each training set (blue) contains one more observation than the previous one, and consequently each test set (red) has one fewer observations than the previous one.*

A variation on this approach focuses on a single forecast horizon for each test set. Figure 4 shows a series of test sets containing only one observation in each case. Then the calculation of accuracy measures is for one-step forecasts, rather than averaging across several forecast horizons.

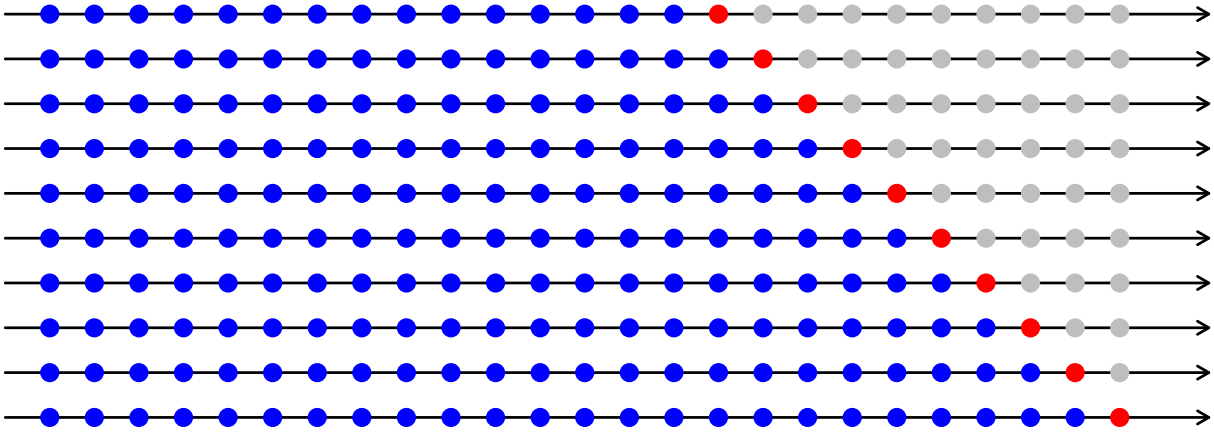

Figure 4: *Time series cross-validation based on one-step forecasts. The blue points are training sets, the red points are test sets and the grey points are ignored.*

In any of these cross-validation approaches, we need a minimum size for the training set because it is often not possible to do any meaningful forecasting if there is not enough data in the training set to estimate our chosen model. The minimum size of the training set depends on the complexity of the model we want to use.

Suppose k observations are required to produce a reliable forecast. Then the process works as follows.

1. Select the observation at time $k+i$ for the test set, and use the observations at times $1, 2, \ldots, k+i-1$ to estimate the forecasting model. Compute the error on the forecast for time $k+i$.

2. Repeat the above step for $i = 1, 2, \ldots, T-k$ where T is the total number of observations.

3. Compute the forecast accuracy measures based on the errors obtained.

This procedure is sometimes known as evaluation on a "rolling forecasting origin" because the "origin" $(k+i-1)$ at which the forecast is based rolls forward in time.

With time series forecasting, one-step forecasts may not be as relevant as multi-step forecasts. In this case, the cross-validation procedure based on a rolling forecasting origin can be modified to allow multi-step errors to be used. Suppose we are interested in models that produce good h-step-ahead forecasts.

1. Select the observation at time $k+h+i-1$ for the test set, and use the observations at times $1, 2, \ldots, k+i-1$ to estimate the forecasting model. Compute the h-step error on the forecast for time $k+h+i-1$.

2. Repeat the above step for $i = 1, 2, \ldots, T-k-h+1$ where T is the total number of observations.

3. Compute the forecast accuracy measures based on the errors obtained.

When $h = 1$, this gives the same procedure as outlined above.

Example: Australian quarterly beer production

To illustrate the above procedure (for one-step forecasts only), we will use the Australian beer data again, with an ARIMA model estimated for each training set. We will select a new ARIMA model at each step using the Hyndman–Khandakar (2006) algorithm, and forecast the first observation that is not in the training data. The minimum size of the training data is set to $k = 16$ observations, and there are $T = 74$ total observations in the data. Therefore, we compute $58 = 74 - 16$ models and their one-step forecasts. The resulting errors are used to compute some accuracy measures:

To calculate the MASE we need to compute the scaling statistic Q, but we do not want the value of Q to change with each training set. One approach is to compute Q using all the available data. Note that Q does not affect the forecasts at all, so this does not violate our rule of not

MAE	11.14
RMSE	14.66
MAPE	2.57%

Table 2: *Error measures calculated on one-step forecasts computed using a time series cross-validation beginning with 16 observations in the training data, and finishing with 73 observations in the training data.*

using the data we are trying to forecast when producing our forecasts. The value of Q using all available data is $Q = 13.57$, so that MASE = 11.14/13.57 = 0.82. This shows that, on average, our forecasting model is giving errors that are about 82% as large as those that would be obtained if we used a seasonal naïve forecast.

Notice that the values of the accuracy measures are worse now than they were before, even though these measures are computed on one-step forecasts and the previous calculations were averaged across 14 forecast horizons. In general, the further ahead you forecast, the less accurate your forecasts should be. On the other hand, it is harder to predict accurately with a smaller training set because there is greater estimation error. Finally, the previous results were on a relatively small test set (only 14 observations) and so they are less reliable than the cross-validation results which are calculated on 58 observations.

4 Conclusions

- Always calculate forecast accuracy measures using test data that was not used when computing the forecasts.
- Use the MAE or RMSE if all your forecasts are on the same scale.
- Use the MAPE if you need to compare forecast accuracy on several series with different scales, unless the data contain zeros or small values, or are not measuring a quantity.
- Use the MASE if you need to compare forecast accuracy on several series with different scales, especially when the MAPE is inappropriate.
- Use time series cross-validation where possible, rather than a simple training/test set split.

References

Hyndman, R. J. and G. Athanasopoulos (2012). *Forecasting: principles and practice*. OTexts. http://otexts.com/fpp.

Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting : the forecast package for R. *Journal of Statistical Software* **26**(3), 1–22.

Hyndman, R. J. and A. B. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**(4), 679–688.