Quantile forecasting with ensembles and combinations

Rob J Hyndman

9 August 2020

Forecasting using possible futures

One way to think about forecasting is that we are describing the possible futures that might occur.

Suppose we are interested in forecasting the total sales in Australian cafés and we train an ETS model and an ARIMA model (Hyndman & Athanasopoulos, 2020) on the data to the end of 2018. Then we can simulate sample paths from these models to obtain many possible "futures". Figure 1 shows the last four years of training data and 5 futures generated from each of the two fitted models.

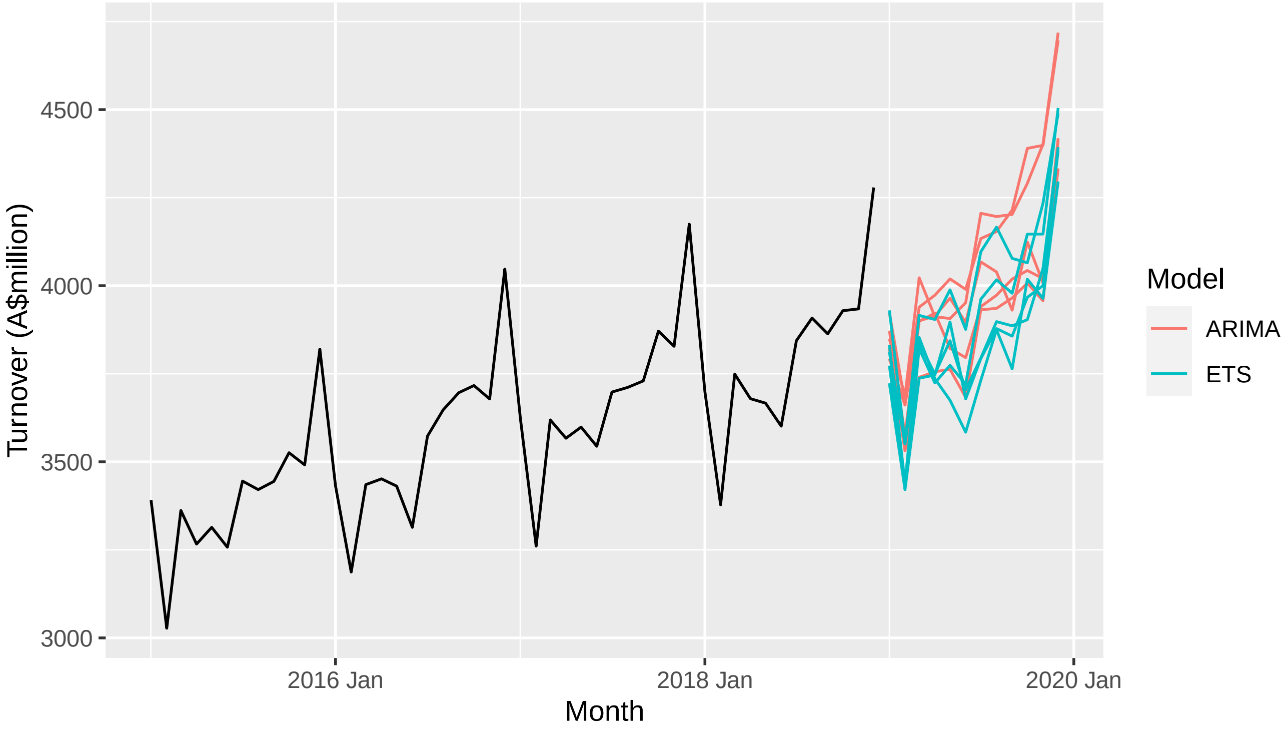

Figure 1: *Future sample paths obtained using an ARIMA model and an ETS model for the Australian monthly café turnover.*

If we repeat this procedure thousands of times for each model, we can obtain a very clear picture of the probability distribution for each future time period. The means of these sample paths are the traditional point forecasts. Traditional 95% prediction intervals are equivalent to finding the middle 95% of the futures at each forecast horizon.

Simulated future sample paths also allow us to answer many more interesting questions. For example, we may wish to find prediction intervals for the total turnover for the next 12 months. This is surprisingly difficult to handle analytically but trivial using simulations — we just need to add up the turnover for each of the simulated sample paths, and then compute the relevant percentiles. We might also want to forecast the maximum turnover in any month over the next year. Again, that is a difficult problem analytically, but very easy using simulations. I expect that simulating future sample paths will play an increasingly important role in forecasting practice because it makes difficult problems relatively easy, and allows us to explore what the future might be like in ways that would otherwise be almost impossible.

Using simulations in forecasting requires a generative statistical model to be used. This is easy using an ARIMA or ETS model, but more difficult if something like a neural network or random forest has been used.

Quantile forecasting

Almost everyone needs probabilistic forecasts whether they realise it or not. Without some kind of probabilistic forecast or other measure of uncertainty, a point forecast is largely useless as there is no way of knowing how wrong it is likely to be. A simple version of a probabilistic forecast is a prediction interval which is intended to cover the true value with a specified probability. Another type of probabilistic forecast is the notion of "safety stock", which is the additional stock to be ordered above the point forecast in order to meet demand with a specified probability.

A more sophisticated way of producing probabilistic forecasts is to generate quantile forecasts. For example, a 90% quantile forecast is a value which should exceed the true observation 90% of the time, and be less than the true value 10% of the time. Median forecasts are equivalent to 50% quantile forecasts. Prediction intervals are often constructed in this way — an 80% prediction interval can be based on the 10% and 90% quantile forecasts. Safety stock can also be computed from quantile forecasts — set the stock order to be the 95% quantile to ensure your probability of being out-of-stock is 5%.

Any statistical forecasting method can be used to produce quantile forecasts by simulation. We simply need to compute the quantiles at each time from the simulated sample paths. Figure 2 shows the deciles for the ETS forecasts (i.e., the 10th, 20th, …, 90th percentiles).

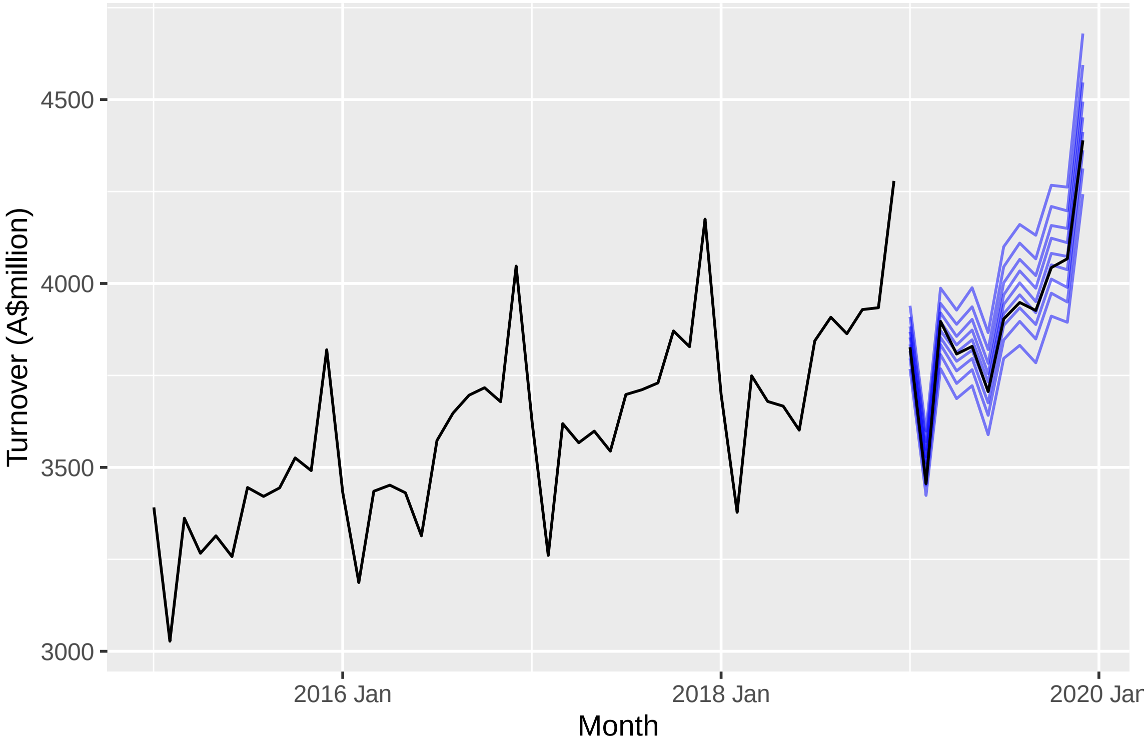

Figure 2: *Blue: Deciles for the ETS forecasts for the Australian monthly café turnover. Black: Observed values.*

Evaluating quantile forecasts

Most business doing forecasting will be familiar with computing accuracy measures for point forecasts such as MAPE or RMSE values. With quantile forecasts, we need to use some alternative measures.

Quantile scores provides a measure of accuracy for each quantile of interest. Suppose we are interested in the quantile forecast with probability p at future time t, and let this be denoted by $f_{p,t}$. That is, we expect the observation at time t to be less than $f_{p,t}$ with probability p. For example, an estimate of the 95th percentile would be $f_{0.95,t}$. If y_t denotes the observation at time t, then the quantile score is

$$Q_{p,t} = \begin{cases} 2(1-p)\big(f_{p,t} - y_t\big), & \text{if } y_t < f_{p,t} \\ 2p\big(y_t - f_{p,t}\big), & \text{if } y_t \geq f_{p,t} \end{cases}$$

This is sometimes called the "pinball loss function" because a graph of it resembles the trajectory of a ball on a pinball table. The multiplier of 2 is often omitted, but including it makes the interpretation a little easier. A low value of Q_p indicates a better estimate of the quantile.

In Figure 2, the 90% quantile forecast for December 2019 is $f_{0.9,t} = 4680$ and the observed value is $y_t = 4389$. Then $Q_{0.9,t} = 2(1-0.9)(4680-4389) = 58$.

The quantile score can be interpreted like an absolute error. In fact, when $p = 0.5$, the quantile score $Q_{0.5,t}$ is the same as the absolute error. For other values of p, the "error" $(y_t - f_{p,t})$ is weighted to take account of how likely it is be positive or negative. If $p > 0.5$, $Q_{p,t}$ gives a heavier penalty when the observation is greater than the estimated quantile than when the observation is less than the estimated quantile. The reverse is true for $p < 0.5$.

Often we are interested in the whole forecasting distribution (not just a few quantiles), and then we can average the quantile scores over all values of p. This gives what is known as the "Continuous Ranked Probability Score" or CRPS (Gneiting & Katzfuss, 2014).

In the Australian café example, we can compute the CRPS values over the 12 months of 2019 for each of the ARIMA and ETS models. To make it more interpretable, we can also compute the CRPS for a simple seasonal naive model, and then we can calculate the "skill score" equal to the percentage improvement for ARIMA and ETS over seasonal naive.

Model	CRPS	Skill score
SNAIVE	68.6	0.0
ARIMA	32.9	52.0
ETS	31.5	54.0

Here, ETS is providing the best quantile forecasts with a skill score of 54.0.

Ensemble forecasting

Ensemble forecasting involves using multiple models and combining the future sample paths to produce the final forecast. If a weighted ensemble is needed, we can make the number of simulations from each model correspond to the required weight.

Ensemble forecasting has been used in weather forecasting for many years, but is not so widespread in other domains. The logic behind ensemble forecasting is that no model is perfect, and the data did not come from a model. As George Box has put it, "all models are wrong, but some are useful" (Box, 1976). Ensembles allow the good features of various models to be included, while reducing the impact of any specific model. It also allows the uncertainty associated with selecting a model to be incorporated into the quantile forecasts.

For the Australian café data, we can combine 10000 simulated sample paths from each of the ETS and ARIMA models, and compute the resulting quantile forecasts from the 20000 sample paths.

Model	CRPS	Skill score
ENSEMBLE	31.4	54.3

The ensemble forecasts are slightly better than either the ETS and ARIMA forecasts in this case. When the component models use very different information, the benefit of using ensemble forecasts is greater.

Combination forecasting

Combination forecasting is a related idea that is more widely used in the general forecasting community. This involves taking a weighted average of the forecasts produced from the component models. Often a simple average is used. For more than 50 years we have known that combination forecasting improves forecast accuracy (Bates & Granger, 1969; Clemen, 1989). One of the reasons for this is that the combination decreases the variance of the forecasts (Hibon & Evgeniou, 2005) by reducing the uncertainty associated with selecting a particular model.

Combinations are almost always used to produce point forecasts, not probabilistic forecasts. A weighted average of several component forecasts gives a point forecast that is identical to taking the mean of the sample paths from the corresponding weighted ensemble.

However, the idea can be used more generally to obtain quantile forecasts as well. Quantiles can not simply be averaged, so we need to take account of the correlations between the forecast errors from the component models when producing quantile forecasts. This is implemented in the `fable` package for R. For the Australian café data, this gives the following result.

Model	CRPS	Skill score
COMBINATION	30.9	54.9

Further improvement has been obtained by taking account of the similarity of the ETS and ARIMA forecasts, rather than simply combining the sample paths as with ensemble forecasting.

Conclusions

We have described several tools for forecasting that are likely to be increasingly used in business forecasting in the future.

- Simulated future sample paths allow us to study how the future might evolve, and allow us to answer more complicated forecasting questions than is possible with analytical methods.
- Quantile forecasts can be produced from these simulated future sample paths and provide a way of quantifying the forecast distributions.
- Quantile scores allow us to evaluate quantile forecasts. Averaging quantile scores gives the CRPS which allows us to evaluate the whole forecast distribution.
- Forecast ensembles combine information from multiple models and often provide a better estimate of future uncertainty than any individual model.
- Forecast combinations are similar to ensembles but also take account of the relationships between the component models. The best forecasts often come from combining models in this way.

Supplements

All the forecasts and calculations produced in this chapter were obtained with the `fable` package for R. The code used is available at ???

References

Bates, JM & CWJ Granger (1969). The combination of forecasts. *Journal of the Operational Research Society* **20**(4), 451–468.

Box, GEP (1976). Science and statistics. *Journal of the American Statistical Association* **71**(356), 791–799.

Clemen, R (1989). Combining forecasts: a review and annotated bibliography with discussion. *International Journal of forecasting* **5**, 559–608.

Gneiting, T & M Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application* **1**(1), 125–151.

Hibon, M & T Evgeniou (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of forecasting* **21**(1), 15–24.

Hyndman, RJ & G Athanasopoulos (2020). *Forecasting: principles and practice*. 3rd edition. Melbourne, Australia: OTexts. OTexts.org/fpp3.